

**Alma Mater Studiorum - Università di Bologna**

**DOTTORATO DI RICERCA IN BIOINGEGNERIA**

**Ciclo XXIX**

**Settore Concorsuale di afferenza:** 09/G2

**Settore Scientifico disciplinare:** ING-INF/06

**COMPUTATIONAL TOOLS AND *IN-SILICO* MODELS  
TO IDENTIFY TRANSCRIPTIONAL  
DETERMINANTS OF CELL PHENOTYPE DECISION  
MAKING**

**Presentata da:** Ing. Marilisa Cortesi

**Coordinatore Dottorato**

Prof. Ing. Elisa Magosso

**Relatore**

Dott. Emanuele D. Giordano

**Controrelatore**

Prof. Ing. Mauro Ursino

**Correlatore**

Dott. Simone Furini

**Esame finale anno 2017**



# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>                             | <b>9</b>  |
| 1.1      | Epithelial-to-Mesenchymal Transition . . . . .  | 10        |
| 1.2      | <i>in-silico</i> Studies of EMT . . . . .       | 13        |
| 1.2.1    | Descriptive Analysis . . . . .                  | 13        |
| 1.2.2    | Computational Models . . . . .                  | 14        |
| <b>2</b> | <b>Model Description</b>                        | <b>17</b> |
| 2.1      | General structure of the model . . . . .        | 17        |
| 2.2      | Boolean model . . . . .                         | 19        |
| 2.2.1    | The model . . . . .                             | 20        |
| 2.2.2    | Attractors' determination . . . . .             | 25        |
| 2.2.3    | Network simplification . . . . .                | 26        |
| 2.2.4    | Edges' determination . . . . .                  | 29        |
| 2.3      | Markov model . . . . .                          | 30        |
| 2.3.1    | Simulation of the model . . . . .               | 32        |
| 2.3.2    | Results . . . . .                               | 33        |
| 2.4      | Discussion . . . . .                            | 37        |
| 2.5      | Materials and Methods . . . . .                 | 40        |
| 2.5.1    | Boolean model . . . . .                         | 40        |
| 2.5.2    | Markov model . . . . .                          | 46        |
| <b>3</b> | <b>Model Validation</b>                         | <b>51</b> |
| 3.1      | Introduction . . . . .                          | 51        |
| 3.1.1    | <i>in-vitro</i> data . . . . .                  | 52        |
| 3.1.2    | <i>in-silico</i> data . . . . .                 | 54        |
| 3.2      | Steady State Analysis . . . . .                 | 54        |
| 3.2.1    | Analysis of the Expected Behaviour . . . . .    | 55        |
| 3.2.2    | Study of the Variation of each Marker . . . . . | 57        |
| 3.3      | Best Time Point Analysis . . . . .              | 58        |
| 3.3.1    | Analysis of the Expected Behaviour . . . . .    | 59        |
| 3.3.2    | Percentage Variation of each Marker . . . . .   | 60        |
| 3.4      | Discussion . . . . .                            | 61        |

|          |   |            |
|----------|---|------------|
| 3.5      | Materials and Methods . . . . .   | 62         |
| 3.5.1    | Experimental data . . . . .   | 62         |
| 3.5.2    | <i>in-silico</i> data . . . . .   | 63         |
| 3.5.3    | Validation . . . . .  | 65         |
| <b>4</b> | <b>Quantification of Protein Markers in Single cells using Optical Microscopy</b> | <b>67</b>  |
| 4.1      | Fluorescence Quantification in Single Bacterial Cells . .                         | 67         |
| 4.1.1    | Background . . . . .  | 67         |
| 4.1.2    | Set-up Calibration . . . . .  | 69         |
| 4.1.3    | Set-up Validation . . . . .   | 75         |
| 4.2      | Fluorescence Quantification in Eukaryotic Cells . . . .                           | 82         |
| 4.2.1    | Background . . . . .  | 82         |
| 4.2.2    | Algorithm description . . . . .   | 85         |
| 4.2.3    | Results . . . . .   | 87         |
| 4.3      | Discussion . . . . .  | 90         |
| 4.4      | Material and Methods . . . . .  | 91         |
| 4.4.1    | Set-up calibration . . . . .  | 91         |
| 4.4.2    | Fluorescence Quantification in Single Bacterial Cells . . . . .                   | 93         |
| 4.4.3    | Fluorescence Quantification in Eukaryotic Cells                                   | 101        |
| <b>5</b> | <b>Cell invasiveness Quantification</b>   | <b>105</b> |
| 5.1      | Cell-Invasiv-O-meter . . . . .  | 105        |
| 5.1.1    | Assay description and aim . . . . .   | 105        |
| 5.1.2    | Algorithm development . . . . .   | 107        |
| 5.1.3    | Results . . . . .   | 109        |
| 5.2      | I-AbACUS (Invasion-Assay Assisted cell CoUnting Software)                         | 110        |
| 5.2.1    | Assay description and aim . . . . .   | 110        |
| 5.2.2    | Description of I-AbACUS . . . . .   | 111        |
| 5.2.3    | Use of the learning algorithm . . . . .   | 114        |
| 5.2.4    | Results . . . . .   | 116        |
| 5.3      | Discussion . . . . .  | 126        |
| 5.4      | Materials and Methods . . . . .   | 127        |
| 5.4.1    | Cell-Invasiv-O-meter . . . . .  | 127        |
| 5.4.2    | I-AbACUS . . . . .  | 130        |
| <b>6</b> | <b>Conclusions and Perspectives</b>   | <b>139</b> |
| 6.1      | Future Developments . . . . .   | 142        |



# Acronyms

**AKT** Protein Kinase B.

**AU** Arbitrary Units.

**CRF** Camera Response Function.

**CSC** Cancer Stem Cells.

**CV** Coefficient of Variation.

**DAPI** 4',6-Diamidino-2-Phenylindole, Dihydrochloride.

**ECM** Extracellular Matrix.

**EF** Empirical Filter.

**EGF** Epidermal Growth Factor.

**ELISA** Enzyme Linked Immunosorbent Assay.

**EMT** Epithelial to Mesenchymal Transition.

**FPKM** Fragments per Kilobase of Transcript per Million Mapped Reads.

**GFP** Green Fluorescent Protein.

**GUI** Graphical User Interface.

**IPTG** Isopropyl  $\beta$ -D-1-thiogalactopyranoside.

**IQR** Interquartile Range.

**Kegg** Kyoto Encyclopedia of Genes and Genome.

**MET** Mesenchymal to Epithelial Transition.

**MSE** Mean Square Error.

**NCBI** National Center for Biotechnology Information.

**OD** Optical Density.

**RNAseq** RNA Sequencing.

**RT-PCR** Real Time Polymerase Chain Reaction.

**SE** Standard Error.

**SHH** Sonic Hedgehog.

**SVM** Support Vector Machine.

**TCGA** The Cancer Genome Atlas.

**TGF $\beta$**  Transforming Growth Factor Beta.

**WNT** Wingless Type.

# Abstract

The study of complex biological processes has significantly benefited from recent technological advancements and the increasing integration between experimental and computational approaches.

In the following this combined approach will be applied to the study of gene expression and specifically to the identification of transcriptional determinants in a phenotypic regulation process.

In particular Chapter 1 introduces the biological phenomenon used as case study, an example of cell fate determination referred to as Epithelial to Mesenchymal transition (EMT) and reviews the computational approaches already used to describe this process.

Chapters 2 and 3, on the other hand, detail the EMT model here presented. The former describes the technical aspects of its development, the analysis that was conducted and its results. The latter presents its validation, that is a comparative analysis of the model's results with *in-vitro* experiments describing the same process.

After the computational study of this complex biological process involved in cell fate determination, a series of software tools are presented. They can be used to analyse experimental data and either inform a computational representation of a process of interest (e.g. parameters identification) or improve the accuracy and reliability of a number of widely used *in-vitro* techniques.

In Chapter 4 the focus is on gene expression quantification at single-cell level, from images acquired with an optical microscope. Two alternative experimental protocols are presented and used to determine the expression of proteins of interest. The ability of these instruments to identify the signal emitted by single-cells makes them able to identify the parameters of the computational model describing the process of interest.

Another set of tools presented in the following (Chapter 5), analyses one of the major consequences of EMT induction in a cell population, i.e. an augmented invasiveness. Two different experimental assays widely used to quantify this characteristic are considered and software tools that improve their reliability and accuracy are developed and

characterized.

Finally in Chapter 6 some conclusive remarks are presented, together with possible future developments of the presented work.

Overall this thesis contributes to the increasingly applied approach that integrates *in-vitro* and *in-silico* techniques when studying biological processes. Indeed it includes a computational representation of a significant example of a phenotype regulation process, built entirely from freely available data and a collection of tools that could be used to analyse experimental data and reliably quantify their results.

# Chapter 1

## Introduction

Cell decision making is a fundamental mechanism in biology, that is critical for the development of organisms from microbes to mammals and a key element for optimal resource utilization and survival in changing environments [1]. This phenomenon relies on a number of complex regulatory networks featuring nested feedback loops [2, 3, 4], and bistable dynamics, blended with gene expression intrinsic noise, i.e. the variability due to stochastic events affecting the limited number of molecules within a single cell [5].

Phenotypic decision making, specifically, is related to the evolution of different sets of observable characteristics and behaviours from genetically identical cells. While being fundamental in embryogenesis and the development of multicellular organisms, this phenomenon is widely exploited: spore forming bacteria use it to survive unfavourable environments, while stoloniferous plants were shown to implement ramet specialization to take advantage of heterogeneous environments [6].

The study of these exceedingly complex biological processes requires the development of innovative approaches, able to measure experimentally the quantities of interest with high precision and at the single-cell level.

One increasingly favoured approach consists in integrating the experimental activity with computational models that are extremely helpful when testing hypotheses regarding the studied phenomenon, since they give access to variables not otherwise quantifiable. Furthermore the *in-silico* analysis can be used to drive the experimental analysis, determining which assays are the most likely to be the most informative.

The interpretation of the experimental data, moreover, could be aided and supported by a more extensive use of computational tools. Indeed they would promote the standardization of protocols and anal-

yses, easing the comparison among results obtained by different laboratories. Furthermore they could be used to integrate information from different sources to develop a more complete representation of the process of interest.

In the following the Epithelial to Mesenchymal transition (EMT) will be used as case study. This is a phenotypic transformation that occurs in mammalian cells and lately has been the object of extensive research being both easy to be induced *in-vitro* and important in a number of pathological processes like cancer progression and fibrosis.

## 1.1 Epithelial-to-Mesenchymal Transition

The EMT is a complex biological process that was firstly observed in the primitive streak of chicken embryos by Elizabeth Hay in the early 1980s. It consists in the transdifferentiation of epithelial cells in mesenchymal ones and is essential in physiological processes like development, wound healing and stem cells differentiation behaviour [7]. Pathological conditions like fibrosis and cancer, exploit this process to their advantage, since fibrogenesis, inflammation and increased cell motility are hallmarks of the EMT [8].

Furthermore it has generally been linked to stemness as it is a fundamental process in the generation of the mesoderm during gastrulation [9], and this association extends to carcinogenesis too, where EMT is involved in the generation of cancer stem cells (CSC), that have self-renewing and tumour-initiating capabilities and are associated with tumour relapse and poor clinical outcome. Even though most of CSC express markers of the mesenchymal phenotype and transcription factors involved in EMT, it has recently been proposed that CSC development and EMT may occur in parallel rather than being causally related [8]. This is partly because CSC are able to repress the transcription factors typical of EMT and undergo a mesenchymal to epithelial transition (MET) upon reaching their suitable metastatic niche [10].

Metastasis formation and cancer dissemination however are key steps in tumorigenesis that have been strictly linked to EMT as it allows cells to dissociate from the primary tumour and intravasate into blood vessels [11]. Yet the role of this phenomenon in cancer progression and patient survival is still debated, due to the complexity of EMT. Indeed this isn't a one-step phenomenon in which epithelial cells suddenly assume the mesenchymal phenotype, but a gradual transformation, composed of many sequential stages, some of them giving origin to intermediate stable phenotypes (Figure 1.1).

Epithelial cells are an important component of many tissues, they

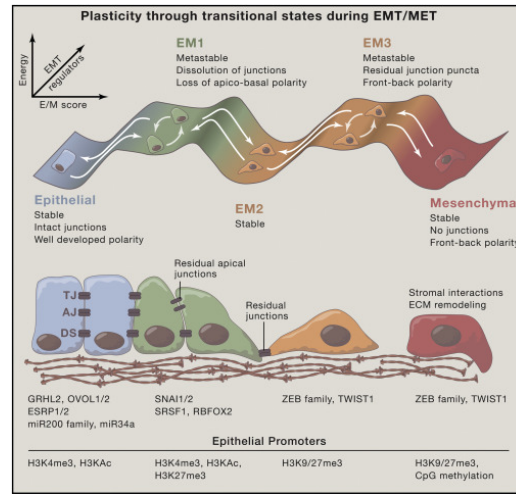


Figure 1.1: Representation of the EMT as a process composed of a number of sequential steps. Figure reproduced from [8].

exhibit apical-basal polarity and interact extensively with each other through specialized cell-cell contact structures such as tight junctions, adherens junctions and desmosomes [12]. These structures dissolve during EMT leading to cells with front-rear polarity, a reorganized cytoskeletal architecture and a modified cell shape (see Figure 1.2 for an example of cells expressing an epithelial, panel **a.** and mesenchymal, panel **b.**, phenotype). These changes reflect also at the molecular level with the downregulation of proteins associated with the epithelial phenotype and the activation of markers associated with the main characteristics of the mesenchymal phenotype: increased cell protrusion and motility, invasive behaviour, ability to degrade the extracellular matrix, resistance to senescence and apoptosis [7] (Figure 1.3).

This process has also gathered significant clinical interest, since EMT has been demonstrated to confer resistance to chemotherapy [15, 16] and immunotherapy [17].

Furthermore MET induction and re-differentiation of mesenchymal cancer cells is an attractive therapeutic option [10, 18, 19, 20] as it provides a way to reuse current therapies reducing both the cost and the time required to get to the clinical application [21]. However several aspects still need to be addressed. One of them is the identification of the phenotype that would grant metastasis abolition and maximal drug sensitivity. There are evidences that a complete EMT reversal might not be ideal and the optimal shift is probably context-dependent and based on the different EMT patterns specific of each cancer [8].

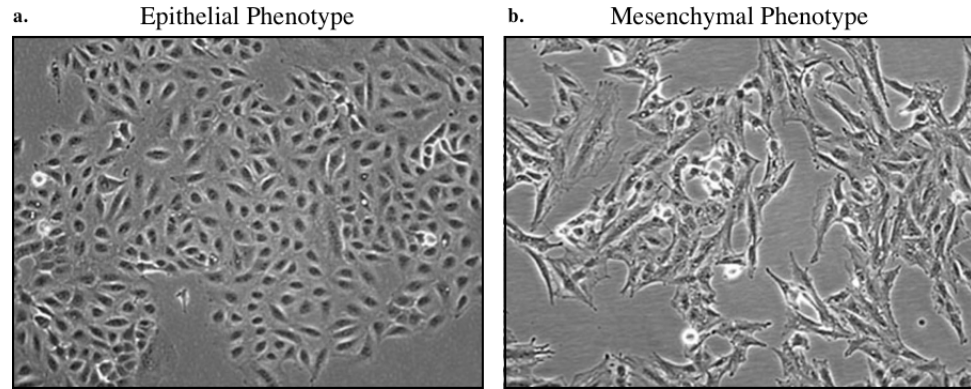


Figure 1.2: Examples of epithelial (a.) and mesenchymal (b.) phenotypes. Images reproduced from [13].

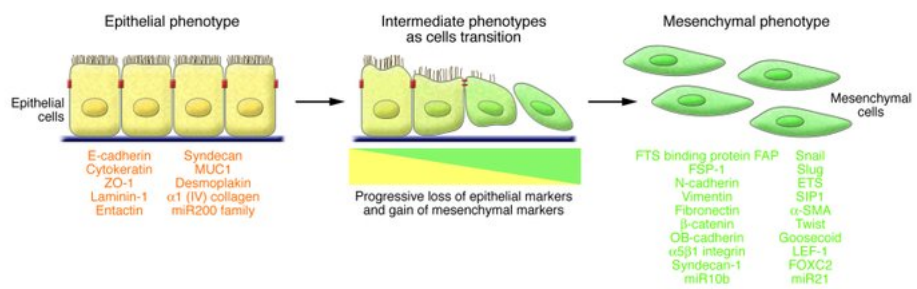


Figure 1.3: Representation of the EMT and of the main markers that characterize its most extreme phenotypes. Figure reproduced from [14].



Furthermore, the induction of EMT reversal might have the side effect of inducing the establishment of secondary metastases, by triggering MET in disseminated cancer cells [10]. Finally most of the markers expressed in the mesenchymal state are implicated in other processes that are fundamental for cell survival, like glucose metabolism and cell cycle, that would be influenced by the therapy in possibly unpredictable ways. As a consequence therapeutic agents that specifically target EMT are not yet available and the role of this phenomenon in both physiological and pathological processes is still debated.

An integrated *in-vitro/in-silico* study of the EMT could be extremely beneficial as it would aid the interpretation of this phenomenon within the multiple contexts in which it manifests, integrating it with other cellular processes, like metabolic changes, that influence or are influenced by this transformation [22, 23, 24, 25].

## 1.2 *in-silico* Studies of EMT

While there are several examples of *in-silico* representation of EMT, differences in their aim make it necessary to distinguish between computational models and tools for the descriptive analysis of this phenomenon. While the former aim to reproduce the EMT and infer the system's behaviour in untested conditions, the latter focus on the analysis of experimental data and have generally a very specific purpose, like the development of a clinically relevant EMT signature.

### 1.2.1 Descriptive Analysis

A recently published example of descriptive computational analysis of the EMT, aims to identify a pan-cancer signature of the mesenchymal phenotype, that could be used in the clinical setting to identify the tumour's phenotype independently of its type [26]. The authors analysed RNAseq data from the online database TCGA [27], to determine which genes had a mRNA level correlated to that of four established EMT markers, called "the seeds". The resulting set of 77 genes was used to evaluate the EMT effect on DNA mutations, gene expression and drug sensitivity. One of the most consequential results of this analysis is the evidence of a strong relation between EMT and immune activation, and especially the direct correlation between mesenchymal scores and increased activation of immune checkpoints, leading to the identification of possible therapeutic targets for mesenchymal tumours, independently of their type.

Other models sought to aid the interpretation of biological data

acquired through the so-called omics technologies. These are experimental techniques recently developed that are characterized by a very high throughput that allows them to analyse the phenomenon of interest on a significantly larger scale. For example they make it possible to quantify the level of expression (either mRNA or protein) of thousands of genes. As one of the main challenges associated with this type of data is their interpretation, a number of computational tools have been developed specifically for this aim. In [28] is presented a model that enables the integration of different omics studies in hybrid interactome networks that allow to identify aberrations in regulatory pathways due to the development of a disease or other physical conditions. As case study, they explore EMT in triple-negative, i.e. lacking estrogen, progesterone and epidermal growth factor receptors, breast cancer cell lines, using both protein expression and phosphorylation data to compare the epithelial and the mesenchymal phenotype. As a result they identify a number of dysregulated modules of the network, whose functionality is strictly connected to focal adhesion and migration.

A similar objective is pursued in [29] where a mixed-effects model is used to study the rewiring of phospho-protein signaling networks due to EMT. They used different combinations of 8 ligands and 5 kinase inhibitors to reveal hidden connections of the network. In this case the results were generated experimentally using ELISA. This work highlighted a significant rewiring of the network due to EMT and identified potential therapeutic targets, that could reverse this transition and restore the epithelial phenotype.

### 1.2.2 Computational Models

Computational models of the EMT, on the other hand, are developed to reproduce the system's behaviour and possibly make inference about it. As an example in [30] a minimalist network composed of 9 genes, is used to represent the induction of EMT with  $TGF\beta$ . The authors used birth-death processes governed by the chemical master equation as mathematical formalism and simulated the model using the Gillespie algorithm [31]. This analysis led the identification of a biphasic EMT dynamic, caused by the sequential action of two bistable switches and regulated by the intensity and duration of the  $TGF\beta$  stimulation.

A more complex EMT representation is described in [32, 33] where the crosstalk between three important EMT pathways ( $TGF\beta$ , SHH, and WNT) is studied. In this case the aim was to evaluate if it was possible to block this transition by removing one or more nodes of the network. This was tested both computationally, identifying the stable motifs that allow to maintain the epithelial phenotype, and experimen-

tally, knocking out up to four nodes at the same time and determining if the combinations suggested by the model were able to abolish EMT completely. This integrated approach reinforced the representation of the studied transition as a multistep process and highlighted how blocking the TGF $\beta$  pathway alone is not sufficient to suppress EMT.

In [22] the metabolic alteration due to EMT is evaluated, through a stoichiometric model of the epidermal growth factor (EGF)-dependent signaling network. This is a scarcely studied, yet important aspect of EMT and metastasis formation, as the profound change in gene expression has significant effects on the metabolism of the cells, that in turn influence gene expression. The framework developed by the authors was able to predict *in-silico* the metabolic phenotype of the considered cells, using gene expression data of the EGF signaling pathway. Specifically they determined that epithelial cells show an higher glycolytic activity, due to the increased flux through the AKT pathway. The most interesting aspect of this work, however, is its generality, as the same approach could be applied to other cell lines, even though the authors point out that the accuracy of the result might be influenced by differences between the experimental models.

Another interesting aspect of EMT is considered in [34], where an agent-based model is used to explore the mechanical changes that occur in the cells during EMT. It was used to describe the interaction forces between different cells and with the extracellular matrix (ECM) during the disruption of the epithelium. Their results show how the repression of one of the main epithelial markers (Cadherin) is fundamental for EMT initiation as it is required for cell-cell junctions dissolution. Furthermore they highlight the importance of the ECM and of the cells spatial distribution for the study of the EMT. Specifically ECM's fiber composition and orientation has been demonstrated to influence cell migration, even though EMT hasn't been specifically analysed. Furthermore the spatial distribution of the cells has a significant effect on their paracrine communication and on the signals that each member of the population receives from the environment. This model would thus be able to explicitly evaluate these aspects and contribute significantly to the study of EMT.

The wide array of formalisms used to describe this process demonstrates its complexity and their minimalist approach the difficulty of realizing a comprehensive yet meaningful computational model. This is because while being able to reproduce the behaviour of interest, these representations generally feature a large number of parameters, that are not easily identifiable as they might not be directly connected to measurable quantities.

Since this level of detail does not allow for a significant expansion of the model, this exploration of the EMT is particularly useful only in the context of specific biological processes such as morphogenesis in early heart development [35] or the progression of pathologies of the eye like proliferative vitreoretinopathy [36].

To address this limitation the EMT is here described using a boolean network. The simplicity of this representation makes it possible to include a large number of functional blocks and study their crosstalk. Furthermore this model does not require the instantiation of any parameter, beside the definition of the network structure and of the update rules.

Another characteristic of the EMT is the significant integration between the behaviour of single cells and that of the population. To include this aspect the boolean network representing single cells was associated with a Markov chain [37], a framework widely used to detail the behaviour of a population.

## Chapter 2

# Model Description

### 2.1 General structure of the model

A tight connection between the behaviour of single cells and that of the population is an important characteristic of the Epithelial to Mesenchymal transition (EMT). Indeed the induction of this phenomenon starts with the transformation of a small number of cells that then contribute to the stimulation of this process in the rest of the population [8].

Explicitly representing this relation is thus both biologically accurate and a relevant aspect when studying EMT induction and progression. In order to be able to integrate all the most important processes involved in the EMT, the single-cell level model is here represented as a boolean network.

This extremely simplified description does not impose strict limitations on the network dimensions and requires only the definition of the network's structure and of the update rules. As a consequence, even processes not entirely characterized experimentally can be considered, since only a minimum amount of information is necessary for the definition of the model.

The population-level model, on the other hand, was represented using a Markov chain. This formalism is widely used for the dynamic simulation of biological processes and can be defined as a network in which each node represents a possible configuration of the system, while edges correspond to the possibility of transitioning between two states. Each connection is characterized by a rate that represents the probability of the corresponding transition. The tight connection between the elements of the model and the quantities that can be measured experimentally have significantly contributed to the success and diffusion of this formalism, that shows a remarkable accuracy even when used

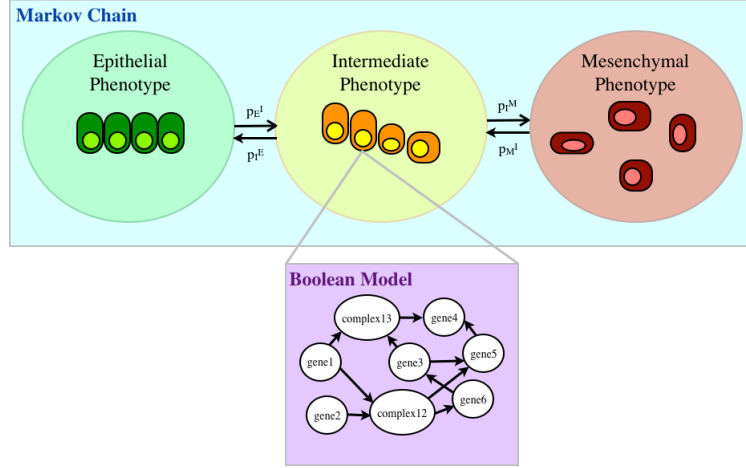


Figure 2.1: Schematic representation of the EMT model. It is composed of a boolean network, describing the behaviour of single cells, coupled with a Markov chain that describes the process of interest at the population level.

to represent complex processes [38].

The integration between the two levels is realized considering that each node of the Markov chain corresponds to a possible phenotype for the cells of the simulated population and that these configurations can be determined as the stable states of the boolean model (Figure 2.1).

Since the boolean network here described was characterized by a large number of fixed point, a signature of genes particularly important for the considered process, was used to combine the attractors and identify the states of the Markov chain.

A similar procedure, further described in the following sections, can be applied to determine the edges and their parameters, thus allowing to completely identify the population level model.

Once built, the model must be validated, that is it must be able to describe the EMT in tested conditions. This step is described on the left end side of Figure 2.2 (red background), where the simulation of the Markov chain in known conditions is compared to experimental data representing the same process.

The validated model could be used to test hypotheses regarding the EMT, simulating its induction in untested conditions or evaluating different mechanisms that might underlie it. These concepts are exemplified in the right end side of Figure 2.2, in the region with the yellow background. The expected behaviour of the system, under the considered assumption, is obtained *in-silico* and then compared to experimental results or integrated with knowledge about the biological process that allows to either determine the hypothesis to be true, or to

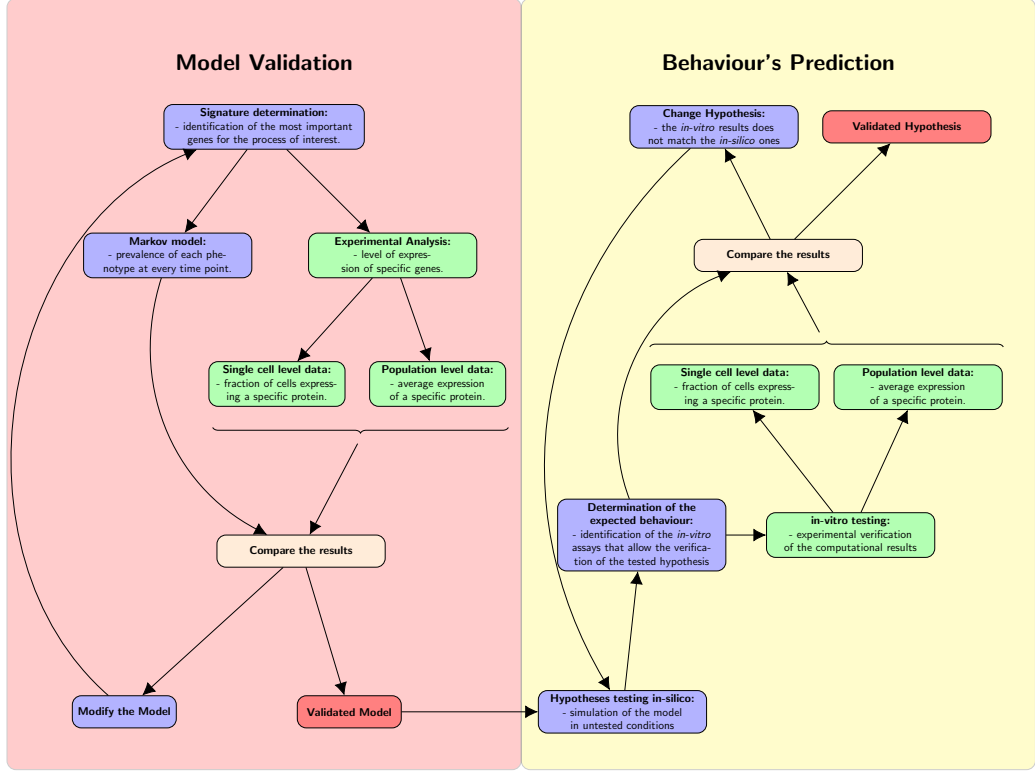


Figure 2.2: Flowchart detailing the validation and the use of the EMT model described in the following.

reject it for its inability to reproduce the behaviour of interest.

## 2.2 Boolean model

A boolean model is a network of fixed topology in which each node represents a protein, or a functional group of proteins, i.e. a complex, while the edges describe the interaction between two nodes. The state of a boolean network comprising  $n$  nodes can be described with a vector, like the one in Equation 2.1, where  $x_i$  represents the value of the  $i$ -th node.

$$V = \{x_1, \dots, x_n\}, x_i \in \{0, 1\} \quad (2.1)$$

Since each protein or complex is simply described as either being present or absent, each node can only assume binary values.

Beside the network's structure, the definition of a boolean model requires the identification, for each node, of an update function (Equation 2.2), that can be used to determine which value to assign to the

current node, given its inputs.

$$x_i(t+1) = f_i(x_{i_1}(t), x_{i_2}(t), \dots, x_{i_{k_i}}(t)), \{i_1, \dots, i_{k_i}\} \subseteq \{1, \dots, n\} \quad (2.2)$$

These functions are applied during the simulation to determine the evolution of the system, described as the set of values assumed by the nodes of the network at each time point, starting from a defined initial condition.

In a boolean network the nodes can be updated either synchronously or asynchronously. In the former case the value of every node is changed at the same time. This protocol makes the network deterministic, since the next state is completely determined by the current one. While being simple to implement this update strategy corresponds to assuming that all the modelled biological processes have comparable dynamics. This assumption is widely accepted as unreasonable for biological processes, that can feature dynamics spanning several orders of magnitude [39, 40, 41, 42, 43].

Thus boolean networks representing gene circuits are generally updated asynchronously. This strategy consists in allowing only one node to change value at every time step. In this case the boolean network becomes a stochastic model, since the next state is determined both by the current one and by the order with which the nodes are updated.

All these aspects will be further detailed in the following, where the computational model describing the EMT is presented.

### 2.2.1 The model

The starting point for the construction of the boolean model was a list of genes widely regarded as involved in EMT initiation and progression. It was obtained from a commercial kit for the study of this process, [44] that, being part of a profiler PCR array, was developed to include markers of all the main processes involved in the EMT (Table 2.1). Beside genes known to be up-regulated or down-regulated during EMT, this list includes genes involved in the production and maintenance of the extracellular matrix and the regulation of the cytoskeleton, that are significantly modified in the first phases of EMT, during the dissolution of the cell-cell contact structures. As a consequence migration and motility markers can be used in the description of the later stages of the studied process, together with genes involved in cell morphogenesis or cell growth and proliferation. Markers of differentiation and development are considered for the importance of the EMT in physiological processes like embryogenesis. A number of transcription factors and genes involved in different signal transduction pathways complete the



|   |  |
|---|--|
| <b>Up-regulated during EMT</b>                          | AHNAK, BMP1, CALD1, CAMK2N1, CDH2 (N-Cadherin), COL1A2, COL3A1, COL5A2, FN1, FOXC2, GNG11, GSC, IGFBP4, ITGA5, ITGAV, MMP2, MMP3, MMP9, MSN, SERPINE1 (PAI-1), SNAI1 (SNAIL), SNAI2, SNAI3, SOX10, SPARC, STEAP1, TCF4, TIMP1, TMEFF1, TMEM132A, TWIST1, VCAN, VIM, VPS13A, WNT5A, WNT5B.    |
| <b>Down-regulated during EMT</b>                        | CAV2, CDH1 (E-Cadherin), DSP, FGFBP1, IL1RN, KRT19, MST1R (RON), NUDT13, OCLN, DESI1, RGS2, SPP1, TFPI2, TSPAN13.  |
| <b>Differentiation and Development</b>                  | AKT1, BMP1, BMP2, BMP7, COL3A1, COL5A2, CTNNB1, DSP, ERBB3, F11R, FOXC2, FZD7, GSC, JAG1, KRT14, MST1R (RON), NODAL, NOTCH1, PTP4A1, SMAD2 (MADH2), SNAI1 (SNAIL), SNAI2, SOX10, TGFB2, TGFB3, TMEFF1, TWIST1, VCAN, WNT11, WNT5A, WNT5B.  |
| <b>Cell Morphogenesis</b>                               | CTNNB1, FOXC2, JAG1, RAC1, SMAD2 (MADH2), SNAI1 (SNAIL), SOX10, TGFB1, TGFB2, TGFB3, TWIST1, WNT11, WNT5A.   |
| <b>Cell Growth and Proliferation</b>                    | AKT1, BMP1, BMP7, CAV2, CTNNB1, EGFR (ERBB1), ERBB3, FGFBP1, FOXC2, IGFBP4, ILK, JAG1, MST1R (RON), NODAL, PDGFRB, TGFB1, TGFB2, TGFB3, TIMP1, VCAN, ZEB1.   |
| <b>Cell Migration and Motility</b>                      | CALD1, CAV2, EGFR (ERBB1), FN1, ITGB1, JAG1, MSN, MST1R (RON), NODAL, PDGFRB, RAC1, STAT3, TGFB1, VIM.   |
| <b>Cytoskeleton Regulators</b>                          | CAV2, KRT7, MAP1B, PLEK2, RAC1, VIM.   |
| <b>Extracellular Matrix and Cell adhesion molecules</b> | BMP1, BMP7, CDH1 (E-Cadherin), CDH2 (N-Cadherin), COL1A2, COL3A1, COL5A2, CTNNB1, DSC2, EGFR (ERBB1), ERBB3, F11R, FN1, FOXC2, ILK, ITGA5, ITGAV, ITGB1, MMP2, MMP3, MMP9, PTK2 (FAK), RAC1, SERPINE1 (PAI-1), SPP1, TGFB1, TGFB2, TIMP1, VCAN.  |
| <b>Signal Transduction</b>                              | CAV2, ESR1 (ER $\alpha$ ), KRT19, TGFB3, AKT1, FZD7, GNG11, RAC1, RGS2, COL3A1, ILK, ITGA5, ITGAV, ITGB1, PTK2 (FAK), FOXC2, JAG1, NOTCH1, EGFR (ERBB1), ERBB3, PDGFRB, RGS2, SPARC, BMP1, BMP2, BMP7, COL3A1, SMAD2 (MADH2), TGFB1, TGFB2, TGFB3, CTNNB1, FZD7, GSK3B, WNT11, WNT5A, WNT5B. |
| <b>Transcription Factors</b>                            | CTNNB1, ESR1 (ER $\alpha$ ), FOXC2, GSC, NOTCH1, GEMIN2, SMAD2 (MADH2), SNAI2, SNAI3, SOX10, STAT3, TCF3, TCF4, TWIST1, ZEB1, ZEB2.  |

Table 2.1: EMT markers used as starting point for the construction of the boolean model. Data reproduced from [44]

list that, in this case, was used to determine the pathways with the highest involvement in the EMT.

The aim of this analysis was to define the structure of the boolean model through the determination of the signaling pathways more relevant for the considered process. In this regard, the online database Kegg (Kyoto Encyclopedia of Genes and Genomes [45]), that integrates genomic and high-order functional information [46] was consulted. At present (23/03/2017) it collects the graphical representations of 509 signal transduction pathways and information on almost  $22 \cdot 10^6$  genes [47], beside almost 2000 disease specific pathways.

A total of 24 pathways, containing at least 6 EMT markers, were isolated (Table 2.2). Most of them are cancer specific or related to the processes mentioned earlier as important for the initiation and progression of the EMT, like focal adhesion and regulation of actin cytoskeleton. Others, like the Hepatitis B pathway, were more surprisingly included, but demonstrate the extension of the changes induced in the human cells by this process.

All the representations of the selected pathways were downloaded from the Kegg website and integrated within a single network in which the interaction attributed to the same gene in different pathways were combined.

As previously described, beside the definition of the network's structure, a boolean model requires an update function for every node. These relations describe the interaction between each node and the ones connected to it, and are thus responsible for the behaviour of the model. In the following three main kinds of relation were considered:

1. activation
2. inhibition
3. complex formation

The first corresponds to a direct proportionality between the expressions of the connected nodes, i.e. when the upstream node is on, it induces the activation of the downstream ones. The inhibiting relation has similar characteristics, but in this case the proportionality is inverse thus the activation of the upstream node cause its targets to switch off.

While these relations focus on the regulation of the gene's activity, the third one describes the functional association of multiple proteins within a complex. For this reason it was modelled with a logical AND (Figure 2.3), that well represents the main requirement for the formation of a complex: the presence of all the proteins that compose it.

Since the same node can be both activated and inhibited these two relations were modelled with the same logical rule, the majority function (Figure 2.4), that states that a node is active if most of its inputs promote its switching on.

The final network, composed of 895 nodes, was analysed to determine the connected components. These can be defined as subgraphs, in which there exists a path that connects each node to every other of the same connected component, while nodes of other subnetworks are unreachable. As an example in Figure 2.5 is reported a network composed of 2 connected components. This step is of great importance, especially for networks of large size, since each subnetwork can be analysed independently due to their complete separation. The connected component analysis of the boolean network describing the EMT identified 171 independent subnetworks whose characteristics are detailed in Table 2.3. Given the substantial difference between the connected

| Pathways   | Number of EMT markers |
|--|-----------------------|
| Proteoglycans in cancer  | 24                    |
| Pathways in cancer   | 23                    |
| Focal Adhesion   | 16                    |
| PI3K-AKT Signaling Pathway,<br>Hippo Signaling Pathway   | 14                    |
| HTLV-I infection   | 12                    |
| Signaling Pathway Regulating<br>Pluripotency of Stem Cells   | 10                    |
| Regulation of Actin Cytoskeleton,<br>Bacterial Invasion of Epithelial Cells,<br>Leukocytes transendotelial migration,<br>Colorectal cancer | 9                     |
| Adherens Junctions, Amoebiasis,<br>Pancreatic Cancer   | 8                     |
| WNT signaling pathway, RAP1<br>signaling pathway, MAPK<br>signaling pathway, FoxO<br>signaling pathway                                     | 7                     |
| Hepatitis B, Chagas Disease,<br>Endometrial cancer, Basal cell<br>carcinoma, Toxoplasmosis, TGF $\beta$<br>signaling pathway               | 6                     |

Table 2.2: Pathways determined to be significantly involved in the EMT, that were used to build the boolean network.

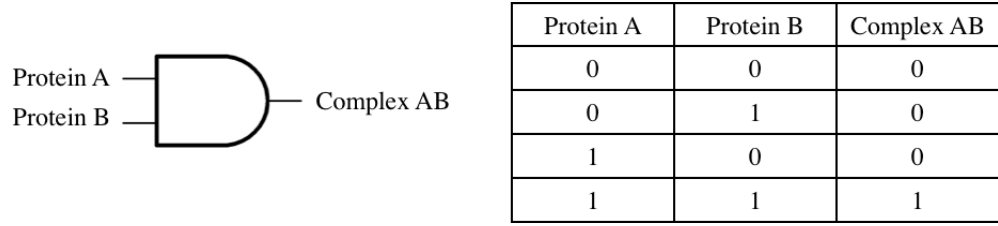


Figure 2.3: The formation of a complex is represented with a logical AND. In this example only two proteins are considered and the update rule (the truth table) states that they both need to be present in order for the Complex AB to form.

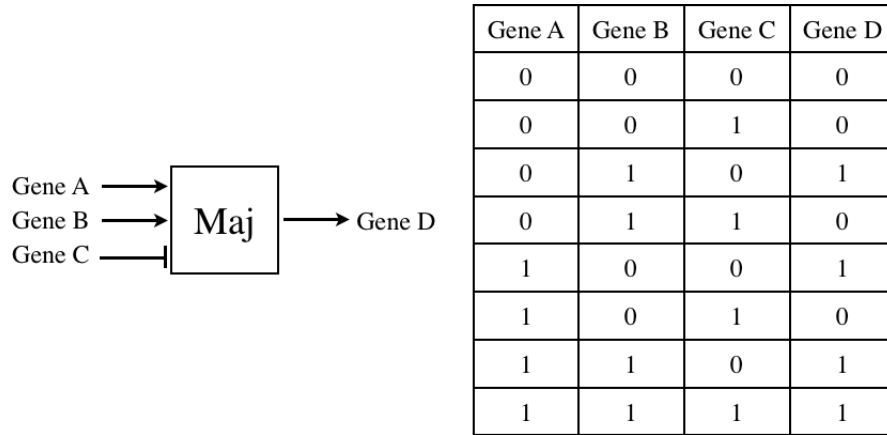


Figure 2.4: Induction and inhibition were both modelled with the majority function. According to this rule a node is on if most of its inputs (at least two in the proposed example) promote its activation.

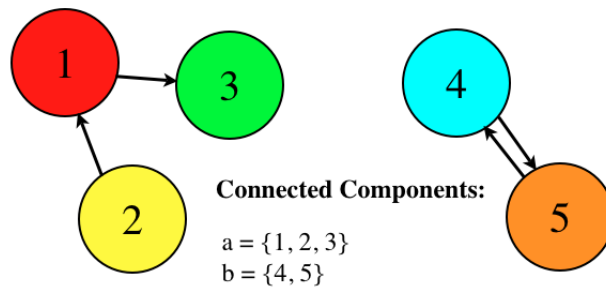


Figure 2.5: Example of network with multiple (2) connected components.

| Number of Nodes | Number of Connected Component |
|-----------------|-------------------------------|
| 1               | 160                           |
| 2               | 3                             |
| 3               | 4                             |
| 4               | 2                             |
| 9               | 1                             |
| 700             | 1                             |

Table 2.3: Result of the connected components analysis performed on the boolean network describing the EMT.

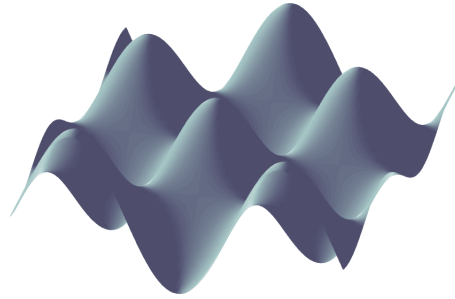


Figure 2.6: A common example used to represent stable states is the potential energy landscape in which the minima represent the attractors of the system.

component with the largest number of nodes, or major component, and the other ones, in the following only the subnetwork composed of 700 nodes will be considered.

### 2.2.2 Attractors' determination

The analysis of the boolean network proceeded with the determination of its steady states. These can be defined as configurations of the system that, once assumed, become permanent, unless a perturbation of sufficient amplitude is applied. A useful metaphor that aids the comprehension of this process consists in representing the state space as a potential landscape (Figure 2.6), where the stable states correspond to the minima. A perturbation able to bring the system beyond a metastable configuration (state 2 in Figure 2.7) leads to the transition between two stable states, identified with the numbers 1 and 3 in Figure 2.7.

While several methods have been developed to determine the at-

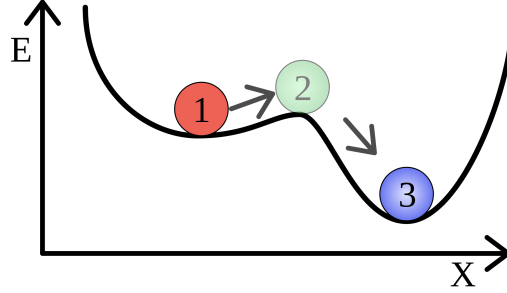


Figure 2.7: Representation of the transition between two stable states (1 and 3). A perturbation of sufficient amplitude forces the system to a configuration associated with a higher potential energy (metastable state 2) and then to a new minimum. Figure reproduced from [48].

tractors of a network [49, 50, 51, 52, 53] the one that is best suited for large systems is its simulation starting from an appropriate number of initial conditions.

The algorithm used in the presented model, simulates the network starting from a random assignment, until all clauses are satisfied, that is all the nodes assume the value determined by their update functions. While unable to identify all the attractors, the repetition of the simulation for an appropriate number of times, starting from different initial conditions, effectively identifies the major stable states of the network, the ones with the largest basin of attraction.

The total number of stable states determined with this analysis is about  $38 \cdot 10^5$ . The large number of stable configurations, in part caused by the non-negligible number of inputs in the network, that remain constant during the simulation, prevent their use as states of the Markov chain.

### 2.2.3 Network simplification

As the dimension of the network and the numerosity of the determined stable states prevents a meaningful analysis of the behaviour of the system, a reduction of the boolean model was operated. A subset of nodes that could recapitulate the main steps of the EMT was determined and then the attractors were condensed, according to the values assumed by these markers.

The EMT signature was identified analysing the boolean network and specifically computing 4 indicators (Table 2.4) that are widely used to determine the importance of each node for the system's functionality.

The number of connections of each node, separated according to

|              |  |
|--------------|--|
| InDeg        | Number of inward connections for each node.  |
| OutDeg       | Number of outward connections for each node. |
| Eccentricity | $\frac{1}{\max(\minPath)}$                   |
| Eigenvalues  | $\lambda_{max}$                              |

Table 2.4: Indicators used to identify the EMT signature. Here minPath represents the minimum distance between each pair of nodes of the network.

their direction, describes the degree of regulation of the corresponding gene (inward connections) and its regulatory power (outward connections). Nodes with high inDeg and/or outDeg are the ones most likely to be included in the signature, since they exhibit an interesting and non-trivial behaviour.

The parameter named eccentricity is defined as the inverse of the longest, shortest path between the considered node and every other one in the network. It gives a measure of the distance between each node and the one that is the most distant in the studied model. Again an high eccentricity is a desirable characteristics as it indicates that the current node is effectively connected to every other element of the network.

The eigenvalues, on the other hand, measure the importance of a node taking into account also the connections of its neighbours. This parameter is thus likely to identify regions of the network that have important regulatory functions.

These 4 indicators were condensed in a score that was used to summarize the importance of each node of the network and determine which ones to include in the signature. To this end, the genes were sorted in decreasing order according to the results of each index. This step was introduced to ensure the equal contribution of the 4 parameters, since the considered measures can assume values that span multiple orders of magnitude and thus indicators with higher numerical values would have been favoured by the simple combination of the results obtained in the previous step. Subsequently the score of each node was calculated as the sum of its rankings.

The nodes with a score below the threshold in Equation 2.3 were considered as part of the signature.

$$th = mean(scores) - 3 \cdot std(scores) \quad (2.3)$$

This choice is justified by Figure 2.8 where the distribution of the scores is reported. Since it is approximately Gaussian, the threshold of Equation 2.3 identifies the 0.01% most significant nodes.

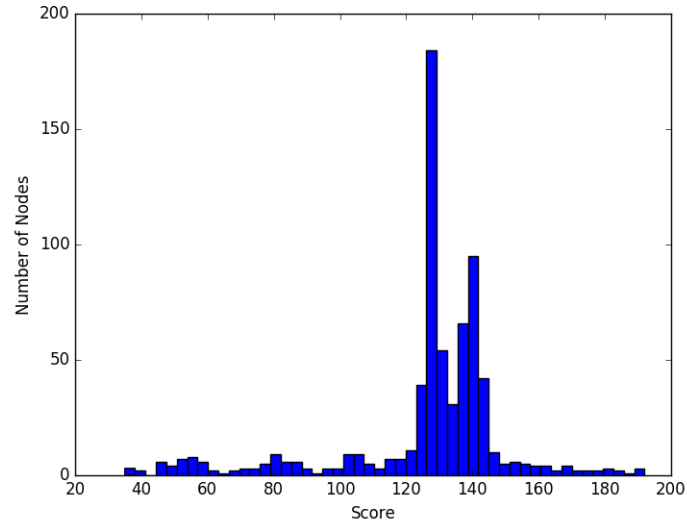


Figure 2.8: Distribution of the scores. Since it is approximately Gaussian the choice of the threshold in Equation 2.3 is justified.

This procedure led to the identification of the genes reported in Table 2.5 that were used to condense the attractors determined through the simulation of the boolean network. This step, exemplified in Figure 2.9, consists in associating the attractors according to the values assumed by nodes that are part of the signature.

| Gene Symbol | Gene Name  | Description  |
|-------------|--|--|
| PCK1        | phosphoenolpyruvate carboxykinase 1                        | Main control point for the regulation of gluconeogenesis.  |
| SAV1        | salvador family WW domain containing protein 1             | Involved in the regulation of protein degradation, transcription, and RNA splicing.                        |
| BRK1        | BRICK1, SCAR/WAVE actin nucleating complex subunit         | -  |
| PTK2B       | protein tyrosine kinase 2 beta                             | Involved in calcium-induced regulation of ion channels and activation of the map kinase signaling pathway. |
| ARAF        | A-Raf proto-oncogene, serine/threonine kinase              | Involved in cell growth and development.   |
| C18995      | P13 protein  | Human T-lymphotropic virus 1.  |
| IKBKE       | inhibitor of nuclear factor kappa B kinase subunit epsilon | Overexpressed in over 30% of breast carcinomas and breast cancer cell lines.                               |



|                        |  |  |
|------------------------|--|--|
| ARHGEF4                | Rho guanine nucleotide exchange factor 4                       | Plays a fundamental role in numerous cellular processes that are initiated by extracellular stimuli that work through G protein coupled receptors. |
| complex-CUL1-RBX1-SKP1 | cullin1 and ring-box 1 and S-phase kinase associated protein 1 | Complex involved in cell cycle progression.  |
| KLK3                   | kallikrein related peptidase 3                                 | Implicated in carcinogenesis.  |
| SSH1                   | slingshot protein phosphatase 1                                | Involved in regulation of actin filament dynamics.   |
| FN1                    | Fibronectin 1  | Involved in cell adhesion and migration processes including embryogenesis, wound healing, blood coagulation, host defense, and metastasis.         |
| ELK1                   | ETS transcription factor                                       | Nuclear target for the ras-raf-MAPK signaling cascade.   |
| ACTB                   | actin $\beta$  | Protein involved in cell motility, structure, and integrity.   |
| C03917                 | Dihydrotestosterone  | -  |

Table 2.5: Signature of genes identified to summarize the EMT. The descriptions here reported were extracted from [54].

The final number of attractors, that were used as states for the Markov chain, was 10639, corresponding to a 97 % reduction with respect to the ones initially determined.


#### 2.2.4 Edges' determination

The determination of the edges of the Markov chain followed a procedure similar to the identification of its nodes. The boolean network was simulated to determine its fixed points, the only difference being in the definition of the initial condition. The attractors determined with the previous analysis were used as starting configurations, after the application of a perturbation, consisting in the random flip of each node of the network with a defined probability.

Initially a 5 % perturbation was applied and the network was simulated as previously described. If the resulting attractor was equal to the starting configuration a perturbation of larger entity was applied, until either a new attractor was reached or the entire configuration was reversed.

The probability of each transition was determined as its frequency, in relation to the total number of transitions exiting the same state. Since the contemporary activation/inhibition of a large number of genes is unlikely, a correction was applied to the values obtained with the frequentist approach. It consists in dividing the transitions probabilities by the distance between the two connected configurations. Thus tran-

|             | Gene 1 | Gene 2 | Gene 3 | Gene 4 | Gene 5 |
|-------------|--------|--------|--------|--------|--------|
| Attractor 1 | 0      | 0      | 0      | 0      | 0      |
| Attractor 2 | 1      | 0      | 0      | 1      | 1      |
| Attractor 3 | 0      | 1      | 1      | 1      | 1      |
| Attractor 4 | 1      | 1      | 1      | 0      | 0      |
| Attractor 5 | 0      | 1      | 0      | 0      | 0      |
| Attractor 6 | 0      | 1      | 1      | 1      | 1      |

|               | Gene 1 | Gene 4 | Gene 5 |
|---------------|--------|--------|--------|
| Attractor 1/5 | 0      | 0      | 0      |
| Attractor 2   | 1      | 1      | 1      |
| Attractor 3/6 | 0      | 1      | 1      |
| Attractor 4   | 1      | 0      | 0      |

Figure 2.9: Exemplification of the attractors condensation step. In this case the signature is composed of Genes 1, 4 and 5 and causes a reduction of about 30 % in the number of attractors.

sitions between similar expression patterns will be favoured, without having to set a defined threshold on the maximum number of nodes that can change at the same time.

## 2.3 Markov model

A Markov chain, is a stochastic model that describes the temporal evolution of a system as the probability of each one of its configurations at every time point. These configurations, or states, are finite in number and are identified by the values assumed by the variables of the system. Another important characteristic of Markov models is that the future state depends only on the current one or, in other terms, the configuration of the system at time  $t$  summarizes the entire simulation up to that point (Equation 2.4).

$$P(X_{\tau+1} = y | X_0 = x_0, X_1 = x_1, \dots, X_n = x) = P(X_{\tau+1} = y | X_n = x) \quad (2.4)$$

This property is verified by a number of systems, provided that all the variables important for the studied behaviour are included in the model, and significantly improves the usability of this framework, by simplifying the simulation.

The definition of a Markov chain consists in identifying the two main elements of the model: the set of its states and the transition matrix. As previously mentioned the states of the model ( $S = \{s_1, s_2, \dots, s_r\}$ ) describe the possible configurations of the system. In the present case, each state represents a phenotype or a pattern of expression of the

genes in the signature of the EMT.

The transition matrix describes how the different states are connected and the probability of each interaction. In the Markov chain formalism the relations between the system's configurations do not change, meaning that this matrix remains constant over time. In the presented approach it was determined using the data obtained with the modified simulation algorithm that allowed the determination of the edges.

In order to simulate the Markov chain, the initial condition, that is the probability of occupancy of each state at  $t=0$ , must be defined. While randomly assigned configurations can be useful to evaluate the range of dynamic behaviours of the model, biologically relevant configurations can be used to test hypotheses regarding the studied system and predict its behaviour in untested conditions.

The initial condition for the EMT model here presented was determined from data retrieved from the Human Protein Atlas [55], an online database that collects information on the expression and localization of most human protein-coding genes products, both at mRNA and protein level. Specifically it contains the characterization of 56 cell lines using deep RNAseq, a technique able to quantify the average level of expression within the population of all human genes. These data are presented in terms of fragments per kilobase of transcript per million mapped reads (FPKM), that is the frequency of a certain sequence, with respect to the total number of recorded fragments, normalized by the length of the corresponding gene. The application of this correction compensates the bias introduced by the presence of coding sequences of different lengths and avoids an overestimation of the level of expression of longer genes.

Beside this quantification, the Human Protein Atlas reports also a qualitative indication of the level of expression of each gene in terms of abundance. This parameter can only assume the 4 values Not Detected, Low, Medium and High, that in the presented model were associated with different ranges of activation of the corresponding gene within the population (Table 2.6). The specific percentage of cells expressing every gene of the EMT signature was then randomly determined respecting the FPKM order, for every level of abundance.

This procedure allows to evaluate the EMT evolution in any one of the cell lines characterized in the Human Protein Atlas. In the following a human lung adenocarcinoma cell line will be considered. This experimental model, denominated A549, is widely considered to show an epithelial phenotype (Figure 2.10) and used as an experimental model for the EMT.

| Abundance    | Expression within the population [%] |
|--------------|--------------------------------------|
| Not Detected | 0                                    |
| Low          | (0,25]                               |
| Medium       | (25,75]                              |
| High         | (75, 100]                            |

Table 2.6: Abundance levels used to summarize the levels of expression in the Human Protein Atlas and the corresponding ranges of activation considered in the presented model.

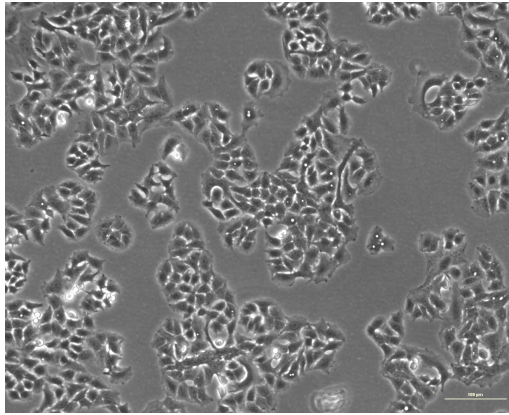


Figure 2.10: Picture of a culture of A549 cells. Their tendency of growing in clusters of tightly connected cells is typical of the epithelial phenotype.

The simulation of an adequate number of *in-silico* cell populations will produce a series of data representing their temporal evolution. The analysis of these data will reveal the most frequently visited states and how their distribution changes over time. Furthermore this framework well adapts to the study of the effect of external inducers, like  $\text{TGF}\beta$ . These simulated experiments, can be used to determine the main effector of a specific behaviour of interest that can then be tested experimentally.

### 2.3.1 Simulation of the model

Multiple simulations of the Markov chain were executed, studying the system's behaviour throughout the entire transition. The results obtained with each simulation, were then combined, leading to an equivalent population of half a million cells, that accurately characterizes the EMT in the considered cell line. This operation, summarized in Figure 2.11, consists in combining the distributions obtained from each simulation, by summing, for each time point, the numbers of cells exhibiting

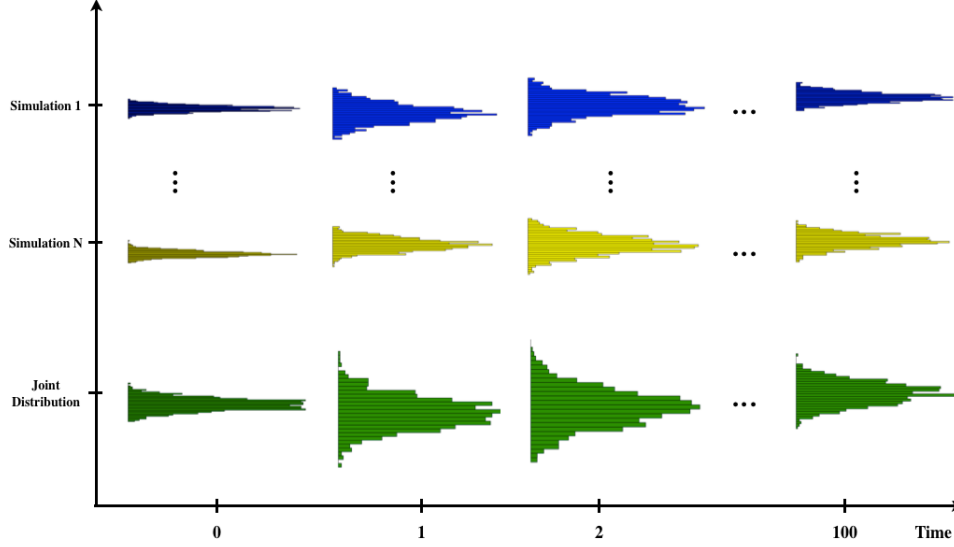


Figure 2.11: Exemplification of the analysis of the results of the Markov model. Each simulation provides a distribution of the possible phenotypes at every time point. The characterization of the EMT, for the tested cell line, is obtained combining the results of all the individual simulations.

the same phenotype.

### 2.3.2 Results

The analysis described in the previous section highlighted the presence both of configurations with high prevalence ( $> 10^5$  cells) and of uncommon patterns of expression ( $< 10$  cells). As a consequence, a representation of the entire phenotype distribution would have masked the more subtle changes in population's composition. This could prevent the correct identification of the first phases of the EMT, that have been demonstrated to occur in only a small number of cells [8].

Thus three different phenotypes distributions were considered, each combining configurations with comparable prevalence. This was achieved considering the highest number of cells recorded during the simulation, for each pattern of expression, and dividing them in three classes:

- High prevalence (over  $10^3$  cells),
- Medium prevalence (between 10 and  $10^3$  cells),
- Low prevalence (below 10 cells).

Furthermore after the first 10 iterations only one distribution every 20 iterations is reported. This is because this experiment was charac-

terized by a fast dynamic, that limited all the changes in population composition within the first few iterations. This is probably determined by the structure of the boolean network in which the flow of information is mainly unidirectional. This characteristic, typical of the signal transduction pathways representations, limits the dynamic of the corresponding network and thus of the Markov chain. This consideration is supported by the distribution of the most prevalent phenotypes ( $> 10^3$  cells), that does not change significantly during the simulation (Figure 2.12).

One notable exception is phenotype 0 that decreases from its initial value. In this pattern of expression the complex-CUL1-RBX1-SKP1 is set to 1. This functional group of proteins is able to repress SMAD2 one of the main regulators of the TGF $\beta$  induced EMT. Thus a decrease in prevalence of the phenotypes expressing this marker is coherent with the induction of the process of interest.

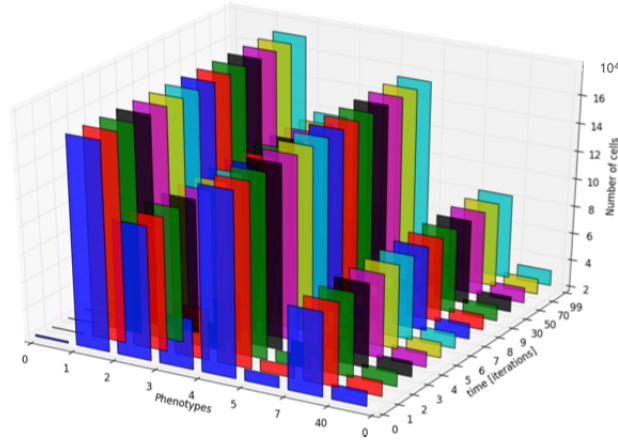


Figure 2.12: Evolution of the most prevalent phenotypes.

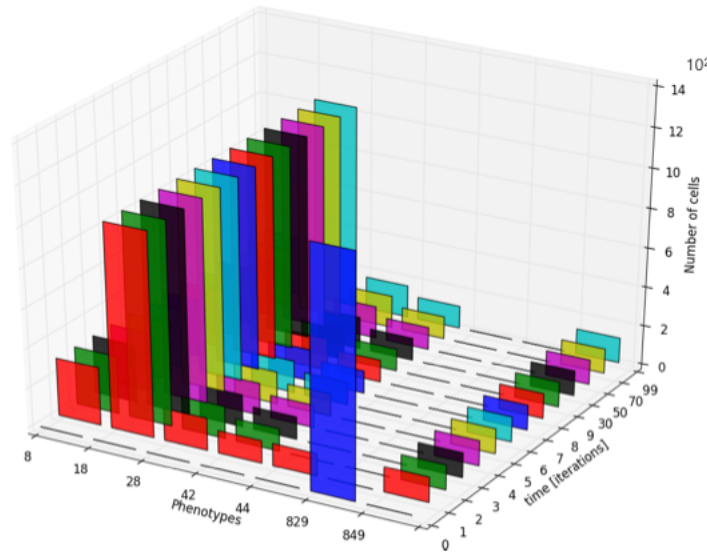
On the other hand, phenotypes characterized by a medium number of cells (between  $10^2$  and  $10^3$ ) show a more interesting behaviour (Figure 2.13). The majority of these configurations (about 70%) increases in prevalence during the simulation, supporting the connection between an increased phenotypic variability and the induction of EMT.

Within the medium prevalence configurations, two other behaviours have been recorded.

Phenotype 829, that in the initial condition is assumed by over 600 cells, rapidly drops to 0. The corresponding configuration, highlighted in red, differs from all the others in Figure 2.13 **b.**, for the expression of the complex-CUL1-RBX1-SKP1. As already remarked the switch-off of this marker is coherent with the first phases of induction of the

TGF $\beta$ -induced EMT.

Phenotype 44, whose configuration is shown in orange in Figure 2.13 **b.**, is characterized by a non-monotonic behaviour. Specifically it is not present in the initial configuration, it appears in the distribution recorded at iteration 1 and then drops again to 0. This oscillation is typical of metastable configurations, in which a small perturbation is sufficient to bring the system to a different state.



|     | PCK1 | SAV1 | BRK1 | PTK2B | ARAF | C18995 | IKBKE | ARHGEF<br>4 | complex-CUL1-<br>RBX1-SKP1 | KLK<br>3 | SSH1 | FN1 | ELK1 | ACTB | C03917 |
|-----|------|------|------|-------|------|--------|-------|-------------|----------------------------|----------|------|-----|------|------|--------|
| 8   | 1    | 1    | 1    | 0     | 0    | 0      | 0     | 0           | 0                          | 1        | 0    | 0   | 0    | 1    | 1      |
| 18  | 1    | 1    | 1    | 1     | 0    | 0      | 0     | 0           | 0                          | 1        | 0    | 0   | 0    | 1    | 1      |
| 28  | 1    | 1    | 1    | 1     | 0    | 0      | 0     | 1           | 0                          | 1        | 0    | 0   | 0    | 1    | 1      |
| 42  | 1    | 1    | 1    | 1     | 0    | 0      | 0     | 0           | 0                          | 1        | 0    | 1   | 0    | 1    | 1      |
| 44  | 1    | 1    | 1    | 1     | 0    | 0      | 0     | 0           | 0                          | 1        | 0    | 0   | 0    | 1    | 0      |
| 829 | 1    | 1    | 1    | 1     | 0    | 0      | 1     | 1           | 1                          | 0        | 1    | 0   | 1    | 1    | 1      |
| 849 | 1    | 1    | 1    | 1     | 0    | 0      | 1     | 1           | 0                          | 0        | 0    | 0   | 0    | 1    | 1      |

Figure 2.13: **a.** Evolution of the phenotypes associated with a number of cells between  $10$  and  $10^3$ . **b.** Table showing the phenotypes corresponding to the distribution in **a.**. Green identifies the configurations that have an increase in prevalence, while red highlight the one that has an opposite behaviour. The configuration that is characterized by a transient behaviour is shown in orange.

The phenotypes with a low prevalence ( $\leq 10$  cells) are the most

numerous (Figure 2.14). Furthermore their distribution experiences a significant increase in standard deviation, since at the beginning of the simulation only one phenotype has been classified in this group.

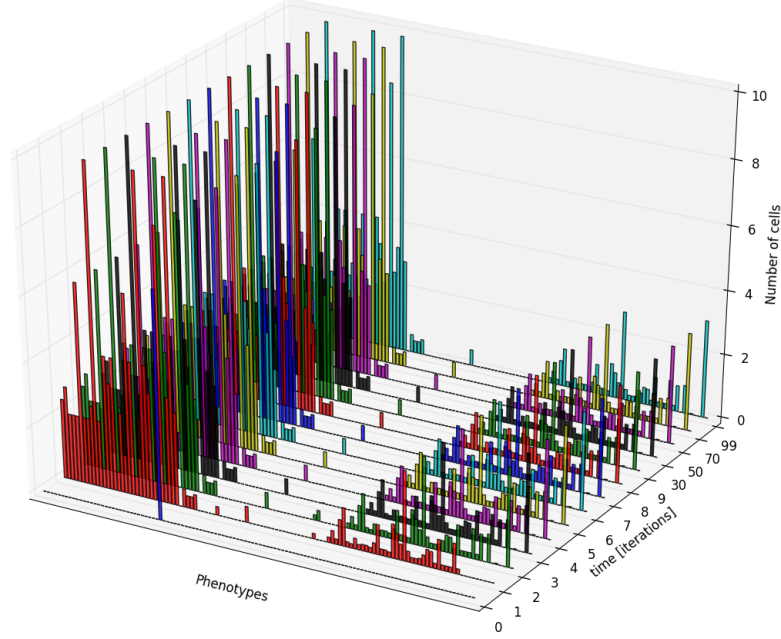


Figure 2.14: Evolution of the phenotypes associated to a number of cells  $\leq 10$ .

Since the configurations of this group are too numerous to be considered independently, only the ones that, at some point during the simulation, had at least 2 cells, were evaluated in detail. These 35 phenotypes, shown in Table 2.7, are mostly characterized by an increase in prevalence. Six configurations, on the other hand, show a transitory behaviour and one, whose interpretation is not straightforward, a decrease in prevalence. All the phenotypes, however share some characteristics. Indeed three markers that should decrease in expression during EMT (complex-CUL1-RBX1-SKP1, SSH1, ELK1) are set to 0, hinting to the possibility that these low prevalence phenotypes might be representative of the patterns of expressions of the sub-population of cells that initiates EMT.

Since the increase in phenotypic variability is an important feature emerging from this analysis, it was further evaluated through the study of the phenotypes with the most pronounced changes in prevalence (over 1% of the corresponding initial value). Indeed a significantly larger number of phenotypes (84) increased in prevalence during the simulation, while only three phenotypes decreased.



The patterns of expression of these 87 phenotypes, were characterized by four markers that were alternatively present in either condition. Specifically ELK1, SSH1 and the complex formed by CUL1, RBX1 and SKP1 were expressed in all the phenotypes that decreased during the simulation. ARAF was set to 0 in all these configurations and was active in about 15 % of the phenotypes that showed an increased prevalence. This behaviour is consistent with the biological function of these markers. ARAF is part of the signal transduction chain that leads to increased cell proliferation, motility and survival, all characteristics typical of mesenchymal cells. SSH1 induces actin stabilization, while ELK1 is implicated in biological processes that induce cell differentiation and apoptosis through the p53 pathway. The decreased expression of both these markers is thus associated with a loss of epithelial characteristics. Furthermore the complex formed by CUL1, RBX1 and SKP1 is a known inhibitor of SMAD2, a transcription factor widely associated with the induction of EMT through  $TGF\beta$ . A reduction in its expression can thus be associated with an increased activity of SMAD2 and the progression of the phenotypic transition.

## 2.4 Discussion

In this chapter a computational representation of a complex biological phenomenon involved in cell decision making is described. This process, called EMT, occurs in epithelial cells and causes them to acquire invasive and migratory capabilities that led to the association between this transformation and metastasis formation [11].

Another important characteristic of EMT is that it encompasses multiple scales. While it occurs at single-cell level, where it is associated to a profound change in both phenotype and behaviour, it has significant repercussions at the population level, as it causes the dissolution of the cell-cell interaction and adhesion structures.

To the author's best knowledge, the model here presented is the first one that describes explicitly this aspect of EMT and has thus the potential to highlight characteristics of this process masked by other representations that focus on a single level of detail or limit their analysis to particular processes within this phenomenon.

EMT in individual cells was modelled with a boolean network describing the signal transduction pathways that were determined to be important for the studied phenomenon. This process, completely automated and operator-independent, was developed to exploit the information collected in freely available databases to extract the representations of the pathways mainly involved in EMT, interpret them and

|            | PCK1 | SAV1 | BRK1 | PTK2B | ARAF | C18995 | IKBKE | ARHGGEF4 | Complex-CUL1-RBX1-SKP1 | KLK3 | SSH1 | FN1 | ELK1 | ACTB | C03917 |
|------------|------|------|------|-------|------|--------|-------|----------|------------------------|------|------|-----|------|------|--------|
| <b>6</b>   | 1    | 1    | 1    | 0     | 1    | 0      | 0     | 1        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>9</b>   | 1    | 1    | 1    | 0     | 0    | 1      | 1     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 0      |
| <b>10</b>  | 1    | 1    | 0    | 1     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 0    | 1      |
| <b>11</b>  | 0    | 1    | 1    | 1     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 0    | 1      |
| <b>12</b>  | 1    | 1    | 1    | 1     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 0    | 0      |
| <b>13</b>  | 1    | 1    | 1    | 0     | 0    | 0      | 0     | 1        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>14</b>  | 1    | 1    | 1    | 1     | 1    | 1      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>15</b>  | 1    | 1    | 0    | 1     | 1    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>16</b>  | 1    | 0    | 1    | 0     | 0    | 0      | 0     | 0        | 0                      | 0    | 0    | 0   | 0    | 1    | 1      |
| <b>17</b>  | 0    | 1    | 1    | 1     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>19</b>  | 0    | 1    | 1    | 1     | 0    | 0      | 1     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>20</b>  | 1    | 0    | 1    | 0     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 1   | 0    | 1    | 1      |
| <b>21</b>  | 1    | 0    | 1    | 0     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>22</b>  | 1    | 1    | 1    | 1     | 0    | 0      | 0     | 0        | 0                      | 0    | 0    | 0   | 0    | 1    | 1      |
| <b>23</b>  | 1    | 1    | 1    | 1     | 1    | 0      | 0     | 1        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>24</b>  | 0    | 1    | 1    | 1     | 0    | 1      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>25</b>  | 0    | 1    | 0    | 1     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 0    | 1      |
| <b>26</b>  | 1    | 0    | 1    | 1     | 0    | 0      | 1     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>27</b>  | 1    | 1    | 1    | 0     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 0      |
| <b>29</b>  | 1    | 1    | 1    | 0     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 1   | 0    | 1    | 1      |
| <b>30</b>  | 0    | 1    | 1    | 1     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 1   | 0    | 1    | 1      |
| <b>31</b>  | 0    | 1    | 1    | 0     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 0    | 1      |
| <b>32</b>  | 1    | 0    | 1    | 1     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>33</b>  | 1    | 0    | 1    | 1     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 1   | 0    | 1    | 1      |
| <b>34</b>  | 1    | 1    | 1    | 1     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 0    | 1      |
| <b>35</b>  | 1    | 1    | 1    | 0     | 0    | 1      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>36</b>  | 0    | 1    | 1    | 0     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 0      |
| <b>37</b>  | 1    | 1    | 1    | 1     | 0    | 0      | 1     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>38</b>  | 0    | 1    | 1    | 0     | 1    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>39</b>  | 1    | 1    | 0    | 1     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>41</b>  | 1    | 1    | 1    | 1     | 0    | 1      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 1    | 1      |
| <b>43</b>  | 0    | 1    | 1    | 0     | 0    | 0      | 0     | 0        | 0                      | 0    | 0    | 1   | 0    | 1    | 1      |
| <b>521</b> | 1    | 1    | 1    | 0     | 0    | 0      | 0     | 0        | 0                      | 1    | 0    | 0   | 0    | 0    | 0      |
| <b>852</b> | 1    | 0    | 1    | 1     | 1    | 0      | 1     | 1        | 0                      | 0    | 0    | 0   | 0    | 1    | 1      |
| <b>859</b> | 1    | 0    | 1    | 0     | 0    | 0      | 1     | 1        | 0                      | 0    | 0    | 0   | 0    | 1    | 1      |

Table 2.7: Table showing the phenotypes with at least 2 cells extracted from Figure 2.14. Green identifies the configurations characterized by an increase in prevalence while the ones in orange show a transitory behaviour. Red marks the phenotype that decreases in prevalence during the simulation.

produce the corresponding boolean network. The simplicity of this representation does not limit significantly the size of the graph that can be analysed and allows to study scarcely known processes, for which kinetic parameters are not available. On the downside, this framework describes each protein as either present or absent, significantly simplifying the representation of gene expression. This might reflect on the model's behaviour and its ability to reproduce *in-vitro* data.

The boolean model describing EMT at single-cell level was simulated, from a large number of randomly determined initial conditions, to determine its main fixed points. These configurations of the network have been linked to the phenotypes that the considered cells can assume and thus the states of a Markov chain describing the phenomenon of interest at population level. Since the configurations retrieved with this approach were too numerous to be used effectively to describe the EMT, they were reduced according to the pattern of expression of a small number of markers, that have been identified as determinant for the behaviour of the boolean network. The connections between these states, and their probability, were then determined, through a procedure similar to the identification of the states of the Markov chain. The only difference was in the definition of the initial condition, that was set to a configuration corresponding to a slightly perturbed stable state for the boolean network. These simulations allowed to determine which fixed points were connected and the frequency of each transition, thus completing the definition of the Markov chain.

The population level model was then simulated, starting from a distribution of phenotypes coherent with a population of lung adenocarcinoma cells, to determine its behaviour during a  $TGF\beta$  induced EMT. Even though the dynamic behavior obtained *in-silico* was faster and less articulate than expected, the results presented in this chapter were coherent with the anticipated ones. Indeed a significant increase in phenotypic variability was recorded, especially when considering phenotypes with low prevalence ( $< 10$  cells). This is concordant with experimental evidences showing that EMT initially occurs in a small number of cells that successively contributes to the induction of this process in the rest of the population [8]. Even the analysis of the phenotypes that experienced the largest variation highlighted signs of the changes induced by EMT at molecular level. Indeed markers associated with actin stabilization, differentiation and apoptosis showed a significant decrease, while a transduction pathway connected to augmented cell proliferation, motility and survival increased in activity. Finally the number of cells expressing a functional complex directly connected to the repression of EMT in the  $TGF\beta$  pathway decreased significantly

during the simulation.

In conclusion the presented model, while showing limited dynamic behaviour, was able to recapitulate some important characteristics of the EMT and, once validated, it might prove useful to study how the molecular changes that occur at single-cell level during the induction of this process influence the behaviour of the population they belong to.

## 2.5 Materials and Methods

### 2.5.1 Boolean model

#### Network Definition

The formalism of the boolean model was used to represent the EMT at the single-cell level. It involves the definition of a graph describing the signal transduction network that regulates the process of interest. Each gene is represented as either being active (node value 1) or switched-off (node value 0) and the relations between different nodes are described with boolean functions.

In the presented model the pathways involved in the EMT were downloaded from the Kegg database [45] and they were chosen for the significant superimposition between the genes that compose them and a list of 84 markers, part of an EMT profiler PCR array [44].

The html files containing the description of the Kegg signal transduction pathways were read using a custom made Python script and combined in a single network, as systematized in Figure 2.15, where two independently downloaded pathways (A and B) share two elements, *gene*<sub>1</sub> shown in red and *gene*<sub>4</sub> represented in green. The combined network maintains all the connections of pathway A but also integrates the additional elements of pathway B: *gene*<sub>5</sub> and *gene*<sub>6</sub> and their connections.

In the presented model, three main relations were considered:

- activation,
- inhibition,
- complex formation.

The first two interactions were described with the majority function, in which a node is considered to be active if most of its inputs determine it to be active. This choice makes it possible to combine directly activating and inhibiting relations interesting the same gene.

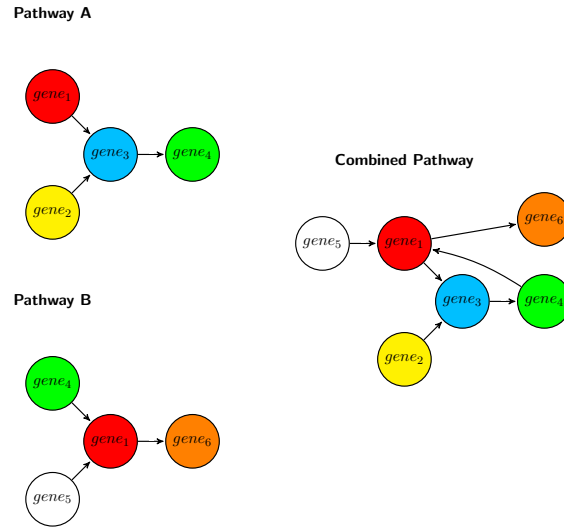


Figure 2.15: Exemplification of the procedure used to combine the pathways downloaded from the Kegg website. When the same gene was present in more than one pathway all its connections were integrated in the final network. In the proposed example the combined pathway integrates all the genes represented in the two networks on the left and their connections.

The formation of a complex, on the other hand, was modelled as a logical AND, to integrate the necessity of the presence of all the nodes that compose it, in order to obtain full functionality.

Successively the connected components were analysed, applying the procedure detailed in Figure 2.16. After selecting a node on the undirected version of the considered graph, a list was created, containing that node and all its neighbours. Successively every node of the list was considered and its neighbours were added to the list. This step was repeated until all the element of that connected component were part of the list. This operation, repeated for all the nodes not included in the already determined subnetworks, allows the determination of all the connected components (Table 2.3). Since the major connected component, the one containing the most elements, is significantly larger than all the other, it will be the only one considered in the following.

### Attractors Determination

The network determined in the previous step was used to compute the stable states of the graph, that correspond to the phenotypes that the modeled cells can assume.

The local search method applied in this phase (Algorithm 1) consists in assigning a random configuration to the nodes of the network

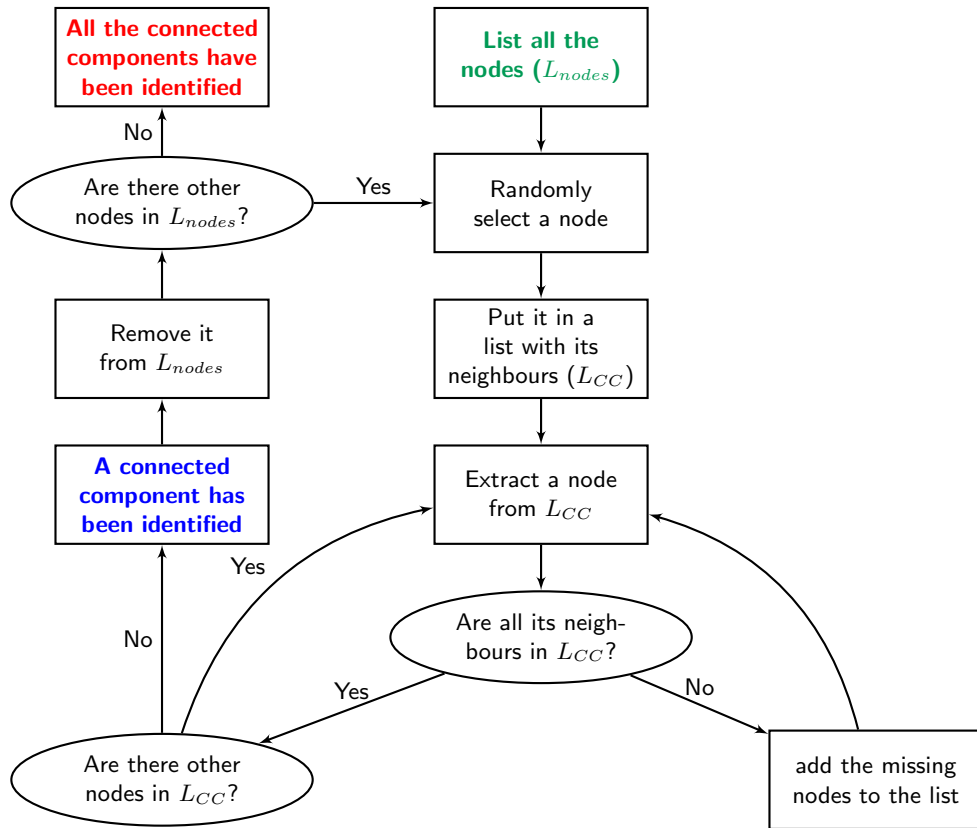


Figure 2.16: Flowchart describing how the connected components of the network were determined. The starting point is highlighted in green.  $L_{nodes}$  represents the list of all the nodes of the network, while  $L_{CC}$  is a variable containing the nodes in the current connected component.

and then updating their values, according to the relations between the nodes defined in the previous step. To improve the accuracy of this representation and model the presence of phenomena with different dynamics, the network is updated asynchronously. Specifically at each iteration, the values of the network's nodes were changed according to a different randomly determined order. The inputs of the graph, nodes that have no incoming connections, will maintain the same value for the entire simulation.

---

**Algorithm 1** Stable States Determination Algorithm

---

```

1: procedure SSD(Network, PercActiveNodes, maxIter)
2:   nodes  $\leftarrow$  number of nodes in the network
3:   initialCondition  $\leftarrow$  defineInitialCondition(nodes, PercActiveNodes)
4:   change  $\leftarrow$  1  $\triangleright$  flag used to track changes in the nodes configuration
5:   iter  $\leftarrow$  0  $\triangleright$  index that keeps track of the number of iterations
6:   currentState  $\leftarrow$  initialCond
7:   while change == 1 — iter < maxIter do
8:     currentState2  $\leftarrow$  currentState
9:     updateOrder  $\leftarrow$  defineUpdateOrder(nodes)
10:    for o in updateOrder do
11:      currentState  $\leftarrow$  updateNode(currentState,o)
12:    end for
13:    iter  $\leftarrow$  iter+1
14:    if currentState==currentState2 then  $\triangleright$  The configuration
      hasn't changed
15:      change  $\leftarrow$  0
16:    end if
17:  end while
18:  return currentState
19: end procedure

```

---

If the result of an iteration is the same sequence obtained in the previous step, the current nodes assignment is considered to be an attractor and saved for further analysis.

This procedure was repeated  $38 \cdot 10^4$  times, starting from randomly determined initial conditions, beside the configurations of network completely active and switched off. Each starting configuration was characterized by a defined probability of activation of the nodes of the network, that was varied between 5% and 95 % with 5% increments. Specifically for each element of the graph a random integer between 0 and 100 was generated and if it was below the activation probability, the corresponding node was set to 1. For each condition  $2 \cdot 10^4$  initial conditions were tested and the simulation continued until either a sta-

ble state or the maximum number of iterations ( $1 \cdot 10^5$ ) was reached (Table 2.8).

|                          |         |
|--------------------------|---------|
| Nodes of the Network     | 700     |
| Initial Conditions       | 380002  |
| Percentage of Activation | 0:5:100 |

Table 2.8: Characteristics of the simulation of the boolean network and resulting number of attractors.

### EMT Signature Determination

Since the attractors determined in the previous step are too numerous to allow a meaningful analysis of the corresponding Markov chain, they were condensed according the pattern of expression exhibited by a subset of nodes determined to be important for the EMT.

This signature was determined considering 4 indicators (Table 2.4) commonly used in network analysis to determine the most important elements of the graph.

Beside the incoming and outgoing connections of each node, the eccentricity was computed. This parameter, defined as the inverse of the longest shortest distance between two nodes, requires the computation of the shortest path between each combination of nodes in the network. In the presented model, the Dijkstra algorithm was applied (Algorithm 2).

It consists in: visiting all the nodes of the graph starting from the current one, and determine which paths connect each pair of nodes more efficiently. In this case the distance between each node and its neighbours is set 1. While this procedure is unable to identify all the shortest paths connecting two nodes, the information obtained with this algorithm, their length, is sufficient to compute the eccentricity.

The last indicator used in this phase are the eigenvalues of the adjacency matrix, that is a matrix describing which nodes of the network are connected. The weights of all the connections was set to 1, and the eigenvalues were computed using the numpy package of the Python language, and specifically its linear algebra section.

These 4 parameters were combined in a score that was computed ordering each node according to the values assumed by the 4 indicators, and then summing the resulting rankings. This choice ensures the equal contribution of each indicator to the final result, as it is independent on the specific range of values assumed by the parameters.

The EMT signature was determined selecting those nodes with a score below a threshold defined as in Equation 2.3. The result of this



---

**Algorithm 2** Dijkstra algorithm [56]

---

```

1: procedure DIJKINSTR(Graph, source)
2:   create vertex set Q
3:   for each vertex v ∈ Graph: do                                     ▷ Initialization
4:     dist[v] ← ∞                                     ▷ Unknown distance from source to v
5:     prev[v] ← undefined                                     ▷ Previous node in optimal path to
      source
6:     add v to Q
7:   end for
8:   dist[source] ← 0                                     ▷ distance from Source to Source
9:   while Q ≠ ∅ do                                     ▷ While the queue is not empty
10:    u ← nodes in Q with min dist[u]                 ▷ Nodes with the least
      distance will be selected first
11:    remove u from Q
12:    for each neighbor v of u do                             ▷ where v is still in Q
13:      alt ← dist[u] + length(u, v)
14:      if alt < dist[v] then                               ▷ A shorter path for v has been found
15:        dist[v] ← alt
16:        prev[v] ← u
17:      end if
18:    end for
19:  end while
20:  return dist[ ], prev[ ]
21: end procedure

```

---

analysis is reported in Table 2.5, where the 15 selected nodes are detailed, together with a short description of their functionalities.

This information, combined with the attractors and the edges previously determined, were used to define a Markov model describing the EMT at population level.

### Edges Determination

The determination of the connections between the attractors of the network was obtained with a modified version (Algorithm 3) of the algorithm described in the previous section. The most important difference is in the definition of the initial condition. Each attractor determined in the previous step was perturbed and then used as starting configuration (10 simulation for every stable state). Initially each node of the stable configuration was flipped with a 5% probability, and then simulated as previously described. If the final configuration is equal to the starting attractor an increased perturbation is applied. This process was applied, increasing the node flipping probability of 5% at each iteration, until either a new attractor is reached or the entire configuration is modified.

These coupled configurations are saved in .txt files that have been used to determine the transition matrix of the Markov chain.

### 2.5.2 Markov model

A Markov chain is identified by a set of states and a transition matrix. They were determined from the results of the boolean network, in particular the set of states was obtained condensing the attractors according to the pattern of expression of the nodes composing the signature, as exemplified in Figure 2.9.

The transition matrix, on the other hand, was obtained determining which states are connected according to the data obtained with the modified boolean network simulation (Algorithm 3). The probability of each transition was quantified using the frequentist approach, that is the number of times a certain transition was recorded, normalized by the total number of transitions exiting from the current state.

The strategy used to compute the attractors does not limit the number of nodes that can be modified at the same time. This is to avoid having to set a defined threshold on this parameter without any specific biological knowledge. However the higher the perturbation, the lower the probability of the corresponding transition. To reflect this consideration, a correction factor was applied to the transition probabilities defined with the frequentist approach. It was defined as

---

**Algorithm 3** Modified Stable States Determination Algorithm

---

```

1: procedure SSD2(Network, attractors, initialPerturbation, maxIter)
2:   nodes  $\leftarrow$  number of nodes in the network
3:   for a in attractors do
4:     perturbation  $\leftarrow$  initialPerturbation
5:     repeat  $\leftarrow$  1  $\triangleright$  flag used to track of the perturbation of the
      nodes configuration
6:     while repeat==1 do
7:       initialCondition  $\leftarrow$  perturbAttractor(attractors[a], pertur-
        bation)
8:       change  $\leftarrow$  1  $\triangleright$  flag used to track changes in the nodes
        configuration
9:       iter  $\leftarrow$  0  $\triangleright$  index that keeps track of the number of
        iterations
10:      currentState  $\leftarrow$  initialCond
11:      while change == 1  $\text{---}$  iter < maxIter do
12:        Simulate the Network as described in Algorithm 1
13:      end while
14:      if currentState==currentState3 then  $\triangleright$  The configuration
        hasn't changed
15:        if currentState  $\neq$  attractors[a] then
16:          repeat=0
17:          return currentState
18:        else
19:          if perturbation < 100 % then
20:            perturbation=perturbation + 5%
21:          else
22:            repeat=0
23:          end if
24:        end if
25:      else
26:        if perturbation < 100 % then
27:          perturbation=perturbation + 5%
28:        else
29:          repeat=0
30:        end if
31:      end if
32:    end while
33:  end for
34: end procedure

```

---

the inverse of the distance between the two configurations and thus favours relations between nodes with similar expression patterns.

This procedure leads to the determination of two fundamental elements for the simulation of a Markov chain, that consists in solving, for every time point, the system of equations in 2.5, where  $s_i(t)$  refers to the fraction of the population exhibiting phenotype  $i$  at time  $t$ , while  $p_{kj}$  is the probability of transitioning from state  $k$  to state  $j$ .

$$\begin{bmatrix} s_1(t+1) \\ s_2(t+1) \\ \vdots \\ s_n(t+1) \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \dots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \dots & p_{2n} \\ & & \dots & & \\ p_{n1} & p_{n2} & p_{n3} & \dots & p_{nn} \end{bmatrix} \cdot \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_n(t) \end{bmatrix} \quad (2.5)$$

The initial condition is the third element required for the simulation of the Markov model, as the system in Equation 2.5 is generally solved using the formulation in 2.6, in which the starting configuration is explicitly shown.

$$\begin{bmatrix} s_1(t+1) \\ s_2(t+1) \\ \vdots \\ s_n(t+1) \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \dots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \dots & p_{2n} \\ & & \dots & & \\ p_{n1} & p_{n2} & p_{n3} & \dots & p_{nn} \end{bmatrix}^t \cdot \begin{bmatrix} s_1(0) \\ s_2(0) \\ \vdots \\ s_n(0) \end{bmatrix} \quad (2.6)$$

In the present application the initial condition was determined from the experimental data collected in the Human Protein Atlas [55]. In this database the level of expression of all the protein-coding sequences of 56 cell lines were quantified using both the FPKM value obtained from RNAseq data and the level of abundance. The latter is a qualitative evaluation of the expression of a specific gene within the population that can only assume 4 values (Not Detected, Low, Medium and High).

For the analysis of the EMT, the lung adenocarcinoma A549 epithelial cell line was considered. The phenotypes initially present within the population were determined assuming the correspondence between the different levels of abundance and the fraction of cells expressing that specific gene (Table 2.6).

The number of cells expressing a certain gene was determined according to the corresponding FPKM value. Specifically, for every abundance level a set of random integers was generated, of the same cardinality as the genes in that expression level. These values were sorted and then assigned to the corresponding element of the signature, according to the corresponding FPKM value.

The expression patterns determined with this analysis were compared to the phenotypes determined with the condensation of the at-

tractors, thus determining the initial prevalence of each state in the Markov chain.

This operation was repeated for the 10 simulations that were executed. The cardinality of every population was  $50 \cdot 10^3$  cells and the simulation was interrupted after 100 iterations (Table 2.9).

|                       |                       |
|-----------------------|-----------------------|
| Initial Population    | $50 \cdot 10^3$ cells |
| Iterations            | 100                   |
| Total Number of Cells | $5 \cdot 10^5$        |

Table 2.9: Specifications of the simulation of the Markov chain.

Since each simulation is independent from the other, all the results were combined to produce a population of  $5 \cdot 10^5$  individuals. This operation, exemplified in Figure 2.11, consists in assigning to each phenotype the prevalence obtained summing those recorded for the individual simulations. For example a phenotype that, at iteration T, is expressed as detailed in Table 2.10 will be assigned a prevalence of 416.

| Simulation | number of cells |
|------------|-----------------|
| 1          | 10              |
| 2          | 32              |
| 3          | 25              |
| 4          | 71              |
| 5          | 56              |
| 6          | 64              |
| 7          | 0               |
| 8          | 100             |
| 9          | 31              |
| 10         | 27              |
| Total      | 416             |

Table 2.10: Example of determination the prevalence of a phenotype (at iteration t). The number of cells expressing each phenotype in the single simulations is summed to obtain the total prevalence.

The combined population was considered to be a more precise representation of the behaviour of interest but the phenotypes were divided according the order of magnitude of their prevalence. This step, executed dividing the considered configurations in three groups according to their maximum number of cells within the simulations, allowed to study with the same accuracy the evolution of phenotypes with significantly different probabilities. Specifically configurations with a maximum prevalence of over  $10^3$  were considered part of the high prob-

ability phenotypes (Figure 2.12), while configurations with at most 10 cells formed the low prevalence group (Figure 2.14). Finally phenotypes with a maximum prevalence between 10 and  $10^3$  formed the third class (Figure 2.13).

The temporal evolution of these configurations was analysed determining which ones were associated to a significant change and linking their pattern of expression to the corresponding biological phenomena. Since in all the considered cases the dynamic behaviour was limited to the first 10 iterations, these were the only ones reported entirely. For all the iterations after the number 9 only one step every 20 was represented.

Successively the phenotypes with the largest percentage variation ( $> 1\%$ ), with respect to the initial value were studied. This operation isolated 87 phenotypes, most of which (84) were associated with an increase in the number of cells expressing them.

The patterns of expression associated with these phenotypes were then considered, to identify possible links between the probability of a phenotype and the induction/progression of the EMT. The fraction of configurations expressing each marker was determined and the phenotypes that experienced an increase of probability were compared to those associated to the opposite behaviour (Table 2.11).

|                        | Increased Prevalence | Decreased Prevalence |
|------------------------|----------------------|----------------------|
| PCK1                   | 81 %                 | 100 %                |
| FN1                    | 18 %                 | 33 %                 |
| ARHGEF4                | 51 %                 | 67 %                 |
| C03917                 | 83 %                 | 100 %                |
| SAV1                   | 75 %                 | 67 %                 |
| BRK1                   | 86 %                 | 67 %                 |
| PTK2B                  | 64 %                 | 100 %                |
| KLK3                   | 55 %                 | 67 %                 |
| ARAF                   | 15 %                 | -                    |
| C18995                 | 24 %                 | 33 %                 |
| IKBKE                  | 2 %                  | 67 %                 |
| ACTB                   | 81 %                 | 100 %                |
| complex-CUL1-RBX1-SKP1 | -                    | 100 %                |
| ELK1                   | -                    | 100 %                |
| SSH1                   | -                    | 100 %                |

Table 2.11: Percentage expression of each marker among the phenotypes that were determined to undergo a significant change in prevalence.

Four of the considered markers were expressed exclusively in one of the two groups and they were determined to have functions associated with the induction of EMT with  $TGF\beta$ .

## Chapter 3

# Model Validation

### 3.1 Introduction

The accuracy and the reliability of a computational model are evaluated via a validation step. This is a comparative analysis between the results of the computational model and those of relevant *in-vitro* experiments. The agreement between the data obtained with the two methods demonstrates the accuracy of the *in-silico* analysis and the reliability of the corresponding results.

The validation process requires the experimental data to be quantitative and at least at the same level of detail as the *in-silico* ones. Specifically a model describing the behaviour of single cells must be compared to experimental data detailed enough to isolate the contribution of each cell to the recorded signal. This is because there exist infinite distributions with the same average and standard deviation, thus population level data could only provide a partial validation.

On the contrary, single-cell level *in-vitro* data could be used to validate population level models, since it is always possible to determine their average value. In this case, however, population data are generally preferred, since they are easier to obtain.

The model validation is generally executed directly comparing the results of the simulations with the experimental ones. In this regard the computational model must be able to quantitatively reproduce the *in-vitro* data. As the output of a large number of experimental assays is in arbitrary units (AU), the analysis is generally conducted comparing the variations from a specific condition, considered as a reference and used to normalize all the available data. This strategy is also applied when the measurement units of the *in-vitro* and *in-silico* datasets differ.

When single-cell level precision is used, probability distributions are often compared, thus allowing to infer information on the stochastic

process that generated the data, through the analysis of the distribution's shape. A simpler framework consists in comparing the moments of the distribution, generally average and standard deviation, regardless of its shape. This approach is commonly used for population level models, but it also applies to single-cell level ones even though it grants only a partial validation, if the data are affected by a high level of measurement noise or if only population level results are available.

When a fully quantitative experimental evaluation of the studied phenomenon is unavailable, the model could be considered to be partially validated if it is able to reproduce qualitatively the trend registered *in-vitro*. This sub-optimal condition can be considered as an intermediate step to guide its further development.

Here, the results obtained simulating the Markov chain were compared to the ones presented in [57], where a population of A549 cells was induced with TGF $\beta$  for 72 hours and the level of expression of a large number of mRNAs was recorded, through Affimetrix microarrays, at 10 time points during the experiment. These data, released through the NCBI's database Gene Expression Omnibus [58] (accession number GSE17708) contain the quantification of the average level of expression of a large number of genes involved in Epithelial to Mesenchymal transition (EMT) induction.

### 3.1.1 *in-vitro* data

The experimental data used to validate the computational model of EMT here described were obtained with an *in-vitro* microarray technique, that allows the quantification of the average level of expression of a large number of genes within the tested population.

These expression levels, available at [58] (accession number GSE17708), were recorded by Sartor et al [57] to test their newly developed mapping tool for gene set enrichment and gene set relation. This system, named ConceptGene, aims to aid the interpretation of high throughput gene expression data determining the functions connected to a list of markers (e.g. differentially expressed genes between two conditions) and identifying connections between Concept types such as biological functions, microRNA target lists, chromosomal regions, or drug target lists.

This analysis allowed them to map the EMT (Figure 3.1) through the determination of the biological processes connected to the changes in gene expression.

While being extremely useful to analyse high throughput data and integrate them with the large base of available knowledge, ConceptGene is not a computational model of biological processes, as it would



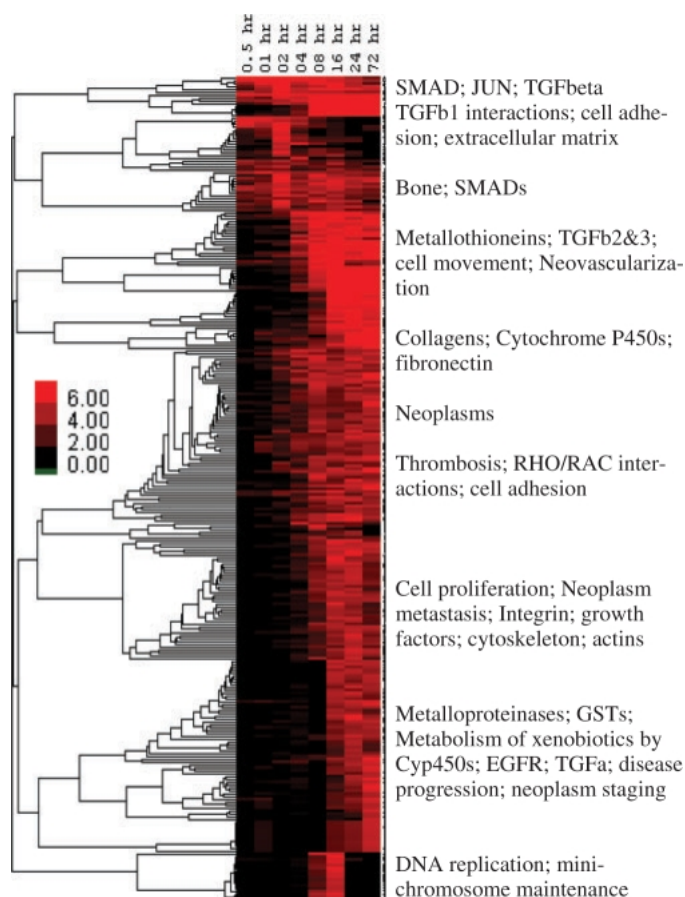


Figure 3.1: Heatmap of enriched Concept profiles throughout TGF- $\beta$ -induced EMT transition in the A549 cell line. All concepts with  $q$ -value  $< 0.05$  for  $\geq 1$  time point were clustered using  $-\log_{10}(\text{p-values})$  to create a “bird’s eye view” of which processes were turning on and off throughout the time course. Figure reproduced from [57].

not be able to infer changes in the EMT induction process caused by untested experimental conditions.

Within the frame of this thesis, the *in-vitro* data presented in [57] were further analysed, to integrate all the probes referring to the same gene. This operation was necessary to allow to compare the *in-vitro* data with the *in-silico* results, that are reported as the fraction of cells expressing each of the tested markers.

Since the measurement unit in which the model results are expressed differs from that of the experimental data, the validation was executed considering all the data normalized with respect to their value at time 0. This operation makes it possible to compare the results of the computational model and the ones presented in [57], assuming that there exists a linear relations between the number of cells expressing a defined marker and its average expression within the population.

### 3.1.2 *in-silico* data

The simulation of the Markov chain describes the EMT results in a set of data displaying the prevalence of each phenotype at every time step. As shown in Figure 2.11 this gives rise to a distribution that evolves over time and can be combined with the ones obtained with the other simulations to generate a complete description of the phenomenon of interest.

While this description allows to determine the most probable EMT induction path and thus the determination of its key steps, the validation of this computational model requires the results to be expressed as average mRNA level for each marker within the population. This transformation was obtained determining the number of cells expressing each gene and assuming a relation of proportionality between this value and the corresponding mean expression.

## 3.2 Steady State Analysis

In the following a partial validation of the presented model will be detailed. In this analysis two elements (C03917 and C18995) will be excluded from the previously determined signature. They correspond, respectively, to dihydrotestosterone and human T-lymphotropic virus 1, that weren't evaluated during the *in-vitro* experiment. This difference, while preventing the complete validation of the model, does not influence the results and the conclusions obtained from the study of the other markers.

The steady state analysis here described consists in two successive comparisons, the first one considers the sign of the variation of each marker with respect to the initial condition, while the second one studies the corresponding amplitude. During the initial analysis, beside the direction of the changes induced by EMT in gene expression, the expected behaviour of each marker was determined, researching the literature and identifying the functionality of each marker within the process of interest.

### 3.2.1 Analysis of the Expected Behaviour

This test was devised to ensure the coherence between the experimental data, the model's results and the expected behaviour. Specifically the difference between the level of expression of each marker at the end of the simulation (or of the experiment) was compared to the one at time 0, determining the sign of the variation. The coherence of this result with the expected behaviour of the network and the literature data was then determined (Table 3.1). In approximately half of the cases

| Marker                     | Sign Variation<br><i>in-vitro</i> | Sign Variation<br><i>in-silico</i> | Biological Process  |
|----------------------------|-----------------------------------|------------------------------------|---|
| SSH1                       | -                                 | -                                  | stabilizes the actin filaments.                           |
| PCK1                       | -                                 | -                                  | regulates glycolysis.                                     |
| complex-CUL1<br>-RBX1-SKP1 | +                                 | -                                  | represses SMAD2.  |
| FN1                        | +                                 | +                                  | increases cell proliferation and motility.                |
| ELK1                       | -                                 | -                                  | involved in cell differentiation and apoptosis.           |
| ARHGEF4                    | -                                 | -                                  | regulates cell-cell adhesion.                             |
| IKBKE                      | -                                 | -                                  | involved in immune response.                              |
| SAV1                       | -                                 | -                                  | regulates apoptosis.                                      |
| BRK1                       | +                                 | -                                  | increases cell motility.                                  |
| PTK2B                      | +                                 | -                                  | required for apoptosis.                                   |
| KLK3                       | +                                 | -                                  | induces cell proliferation.                               |
| ARAF                       | -                                 | +                                  | induces angiogenesis, cell proliferation and growth.      |
| ACTB                       | +                                 | -                                  | maintains adherens junctions and is involved in invasion. |

Table 3.1: Sign of the variation of the level of expression of the signature's markers at steady state and comparison with the expected behaviour. The latter is color coded, where green is associated with an expected increase in expression while red marks the proteins anticipated to decrease in concentration during the EMT. Orange was used to identify one gene whose expected behaviour was not univocally determined.

(46%) the three methods are concordant. In particular SSH1, a protein involved in the process of actin stabilization [59], is determined to decrease during the EMT. This is coherent with the dissolution of cell-cell interaction structures and the reorganization of the cytoskeleton.

Similarly the level of PCK1 at the end of the simulation, results lower than the initial one. This enzyme is involved in the regulation of glycolysis that in [22] is demonstrated to be higher for epithelial cells.

Fibronectin (FN1), on the other hand, increases in both the simulated and the experimental data. This is coherent with the function of this protein that is involved in the regulation of focal adhesion such that its specific expression increases cell proliferation and motility [60].

ELK1, a protein involved in cell differentiation and apoptosis through the p53 pathway [61], is another marker coherently determined to decrease.

The same behaviour is recorded for ARHGEF4, an important element of the cadherin mediated cell-cell adhesion [62].

SAV1, a tumour suppressor involved in apoptosis induction [63], is determined to decrease by all the considered methods.

The complex formed by CUL1, RBX1 and SKP1, on the other hand, is determined to decrease *in-silico*, while a mild increase was recorded *in-vitro*. This complex is known to repress SMAD2 one of the main transcription factors that drive EMT [64] and thus it should be expected to decrease, as predicted by the model. This inconsistency might be caused by the method used to determine the level of expression of the complex from that of the single genes. In the experimental data used in this analysis, each component of the complex was measured independently and the lowest concentration was considered to be representative of the one of the complex, to reflect the impossibility of forming the functional group if one or more elements are missing. This might lead to an over-estimation of the complex's level, as it assumes that all the proteins of the lowest expressed gene are bound to the other elements of this functional group.

IKBKE is a kinase that has been demonstrated to be involved in immune response in breast cancer [65]. Since an activation of the immune system is tightly connected to cancer progression, this marker is supposed to increase, while both the simulations and the *in-vitro* data register its decrease. This might be caused by differences in the experimental model. Indeed EMT might follow alternative paths in breast and lung cancer, thus accounting for this discrepancy.

BRK1 induction is associated with an increase in cell motility [66] and thus its expression is expected to increase during EMT. The computational model is unable to reproduce this result, probably due to the effect of additional regulation of this marker not included in the *in-silico* representation.

PTK2B is a tyrosine kinase required for apoptosis [67] that is thus expected to decrease, as it happens with the computational model. The otherwise increased expression recorded *in-vitro* can be attributed to the involvement of this marker in cell invasion [68], or to alternative splicing that leads to the different isoforms of this protein that might

not be recognized with the same efficiency experimentally.

KLK3 has been demonstrated to induce cell proliferation [69], thus EMT induction should promote its expression. The microarray experiment is able to capture this behaviour while the model displays a slight decrease in the number of cells that produce this protein. This might be caused, as already remarked for BRK1, by additional regulatory pathways acting on this gene but out of the knowledge included by the model.

ARAF is a member of the signal pathway that leads to angiogenesis and augmented cell proliferation and growth [70]. It is thus expected to increase, as shown by the computational model. However alternative splicing variants of these protein have been recorded, that might be associated with a higher variability between the probes evaluating the expression of this marker and thus to a lower precision of their combined value.

Finally actin- $\beta$  (ACTB) is expected to decrease, for its involvement in the maintenance of adherens junctions [71], but other sources determine it to increase [72]. These contrasting evidences suggest that the behaviour of this marker, often used as a housekeeping gene, might be more complex than normally considered. This might be associated with an obsolete version of the signalling pathway in the Kegg database that would, in turn, influence the *in-silico* results.

### 3.2.2 Study of the Variation of each Marker

In addition to the study of the sign of the variation of each marker its amplitude was also considered. Overall, the magnitude of the variation was comparable between the two methods in 54% of the cases. Specifically the differences between the level of expression at the end of the experiment/simulation and its starting value was computed for each marker and ranked in decreasing order. In seven out of thirteen cases (Table 3.2) the same gene was ranked similarly with both methods (rank difference less than the average). The mean difference was determined to be  $4.69 \pm 2.84$  and it was approximated to the nearest integer (5) for the definition of the threshold used to isolate the markers ranked similarly *in-silico* and *in-vitro*.

These results determined the steady-state analysis to be inconclusive. Indeed both the percentage of markers that were determined to vary in the same direction and those whose variation was ranked similarly, were close to 50% (that is the result to be expected when comparing two independent variables). This outcome might be caused by an erroneous assumption regarding the equivalence between the results at the end of the simulation and those obtained *in-vitro* after 72

| Marker                 | <i>in-vitro</i> Ranking | <i>in-silico</i> Ranking | Difference |
|------------------------|-------------------------|--------------------------|------------|
| FN1                    | 1                       | 5                        | 4          |
| PCK1                   | 2                       | 9                        | 7          |
| ARAF                   | 3                       | 10                       | 7          |
| SSH1                   | 4                       | 1                        | 3          |
| PTK2B                  | 5                       | 3                        | 2          |
| ARHGEF4                | 6                       | 2                        | 4          |
| ELK1                   | 7                       | 1                        | 6          |
| SAV1                   | 8                       | 7                        | 1          |
| ACTB                   | 9                       | 4                        | 5          |
| IKBKE                  | 10                      | 11                       | 1          |
| KLK3                   | 11                      | 6                        | 5          |
| BRK1                   | 12                      | 8                        | 4          |
| complex-CUL1-RBX1-SKP1 | 13                      | 1                        | 12         |

Table 3.2: Rankings of markers variations.

hours of induction with TGF $\beta$ . Indeed, while both the simulation and the expression level of the considered markers had reached a plateau at the end of the respective experiments, a conversion factor between iterations and hours was not determined.

### 3.3 Best Time Point Analysis

To test the plausibility of this hypothesis and draw a meaningful comparison between the *in-silico* results and the *in-vitro* data, a correlation analysis was performed. The aim of this procedure was to match the phenotypes distribution at the end of the simulation to the experimental results that shows the best agreement.

For each marker it consisted in comparing the variation from the starting condition obtained at the end of the simulation with the same quantity evaluated at all the timepoints tested experimentally. The result of this analysis, reported in Table 3.3, determined that the results obtained with the computational model show the best agreement with the experimental values obtained 30 minutes after the beginning of the induction with TGF $\beta$ .

This result, combined with the limited and fast dynamic observed in Figures 2.12, 2.13, 2.14 suggests that an important regulatory element for EMT progression might not be represented in the model. This omission, in turn, causes this transition to stall and prevent it from proceeding beyond its first phases.

To test if the higher correlation registered with the experimental data at T=0.5 h reflects in a better concordance between the experimental data and the simulation's results, the analysis of the expected

| Time [h] | r    |
|----------|------|
| 0.5      | 0.75 |
| 1        | 0.48 |
| 2        | 0.66 |
| 4        | 0.51 |
| 8        | 0.48 |
| 16       | 0.58 |
| 24       | 0.56 |
| 72       | 0.56 |

Table 3.3: Correlation between the steady state variation, with respect to the initial condition, of the simulated data and the experimental results at the different time points tested.

behaviour and that of the percentage variation of each marker were repeated, considering the experimental data recorded 30 minutes after the induction with  $TGF\beta$ .

### 3.3.1 Analysis of the Expected Behaviour

This analysis was conducted as described in the previous section. The sign of the variation between the number of cells expressing each marker at steady state and at the beginning of the simulation was compared to the same quantity computed on the data recorded *in-vitro* after 30 minutes of induction with  $TGF\beta$ .

The results, reported in Table 3.4, show a remarkable improvement with respect to the ones in Table 3.1. Now the agreement between

| Marker                     | Sign Variation<br><i>in-vitro</i> | Sign Variation<br><i>in-silico</i> | Biological Process  |
|----------------------------|-----------------------------------|------------------------------------|---|
| SSH1                       | -                                 | -                                  | stabilizes the actin filaments.                           |
| PCK1                       | -                                 | -                                  | regulates glycolysis.                                     |
| complex-CUL1<br>-RBX1-SKP1 | -                                 | -                                  | represses SMAD2.  |
| FN1                        | +                                 | +                                  | increases cell proliferation and motility.                |
| ELK1                       | -                                 | -                                  | involved in cell differentiation and apoptosis.           |
| ARHGEF4                    | -                                 | -                                  | regulates cell-cell adhesion.                             |
| IKBKE                      | -                                 | -                                  | involved in immune response.                              |
| SAV1                       | -                                 | -                                  | regulates apoptosis.                                      |
| BRK1                       | +                                 | -                                  | increases cell motility.                                  |
| PTK2B                      | -                                 | -                                  | required for apoptosis.                                   |
| KLK3                       | +                                 | -                                  | induces cell proliferation.                               |
| ARAF                       | +                                 | +                                  | induces angiogenesis, cell proliferation and growth.      |
| ACTB                       | -                                 | -                                  | maintains adherens junctions and is involved in invasion. |

Table 3.4: Expected behaviour analysis obtained using the microarray data recorded 30 minutes after induction. Green is associated with an expected increase in expression while red marks the proteins anticipated to decrease in concentration during the EMT. Orange was used to identify one gene whose expected behaviour was not univocally determined.

*in-vitro* data, simulation results and expected behaviour is achieved 77 % of the times. Only three markers, IKBKE, BRK1 and KLK3, show alternative behaviours. Specifically both KLK3 and BRK1 are determined to decrease by the model, while their experimental and expected behaviour show an increase. This discrepancy, recorded also during the steady state analysis, supports the idea that an important regulatory element of the TGF $\beta$ -induced EMT might not have been included in the model, causing both the transition to stall and these markers to decrease.

IKBKE, on the other hand, showed a decrease both *in-silico* and *in-vitro*, while it was expected to increase due to its connection to the immune response that activates during breast cancer. As for the previous analysis, this inconsistency might be attributed to differences in the experimental model that might cause the EMT to proceed alternatively in breast and lung cancer. Under this hypothesis, the percentage of the marker whose behaviours is correctly reproduced *in-silico* rises to 85 %.

### 3.3.2 Percentage Variation of each Marker

The analysis proceeded with the evaluation of the amplitude variations. The difference between the number of cells expressing each marker at the end of the simulation and the corresponding value set in the initial condition, was computed and ranked in decreasing order. The same procedure was applied to the *in-vitro* data recorded 30 minutes after the beginning of the TGF $\beta$  induction.

In this case, 10 out of 13 markers had a difference in ranking below 5, with an improvement of over 20% with respect to the steady state analysis. These data, reported in Table 3.5 and Figure 3.2, show a better agreement between the experimental data and the *in-silico* results, with a decrease in the mean ranking difference from  $4.69 \pm 2.84$  to  $3.5 \pm 2.06$ . In particular their graphical representation highlights a very good agreement on most of the data, with only three markers deviating significantly from the trend identified by the linear fitting (Figure 3.2 **a.**,  $r=0.64$ ). Indeed the exclusion of these three points, corresponding to PCK1, BRK1 and ARAF leads to an improvement in correlation of about 30% (Figure 3.2 **b.**,  $r=0.88$ ).

Furthermore the two ranking most frequently assigned, 2 and 4, were both below the threshold defined during the steady state analysis and less or equal to the average ranking difference obtained for T=0.5 h (approximated to the nearest integer).



| Marker                 | <i>in-vitro</i> Ranking | <i>in-silico</i> Ranking | Difference |
|------------------------|-------------------------|--------------------------|------------|
| PCK1                   | 1                       | 8                        | 7          |
| ELK1                   | 2                       | 1                        | 1          |
| complex-CUL1-RBX1-SKP1 | 3                       | 1                        | 2          |
| ARHGEF4                | 4                       | 2                        | 2          |
| SSH1                   | 5                       | 1                        | 4          |
| BRK1                   | 6                       | 8                        | 2          |
| PTK2B                  | 7                       | 3                        | 4          |
| FN1                    | 8                       | 5                        | 3          |
| ARAF                   | 9                       | 10                       | 1          |
| SAV1                   | 10                      | 7                        | 3          |
| KLK3                   | 11                      | 6                        | 5          |
| ACTB                   | 12                      | 4                        | 8          |
| IKBKE                  | 13                      | 9                        | 4          |

Table 3.5: Rankings of the markers variations considering the experimental data recorded 30 minutes after induction with TGF- $\beta$ .

### 3.4 Discussion

In this chapter a partial validation of the computational model of EMT is described. Experimental data acquired using the microarray technique and describing the induction of the EMT with TGF $\beta$ , were compared to the results obtained with the model.

Two main comparisons were drawn, the first one studied the sign of the variation in the level of expression of each marker. This analysis, repeated both at steady state and at the time point that was determined to be associated with the highest correlation between *in-silico* and *in-vitro* results, studied the coherence among the two considered methods and the expected result.

The second comparison detailed in this chapter, analysed the relative amplitude of the variation of each marker. This study aimed to determine if the modifications induced by EMT on the considered genes were comparable *in-silico* and *in-vitro*.

Taken together, these results demonstrate that this EMT model well represents the first phases of the considered transformation, with the highest correlation between *in-silico* and *in-vitro* data recorded at 30 minutes after induction. The simulated transition, however, was apparently stalling after this initial step, suggesting the possibility that an important regulator of EMT might have been excluded from the model.

A more detailed analysis of the function of two elements of the considered signature (BRK1, KLK3), whose variation was not correctly reproduced by the model, could aid the identification of the regulation elements that, once introduced in the *in-silico* representation, would

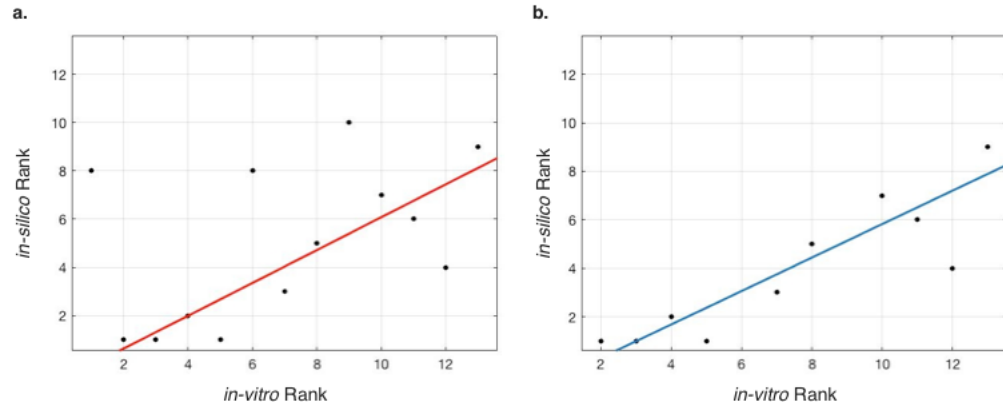


Figure 3.2: Scatter plot showing the correlation between the markers ranking *in-vitro* and *in-silico*. In **a.** all the markers are considered, while in **b.** the rankings of the three outliers were removed.

hopefully allow the computational study of the entire phenotypic transition.

Despite some limitation, the presented model, due to its novel approach and the immediate integration with experimental data holds a significant potential for development and is expected to prove useful for the study of EMT and cell phenotypic transformations in general.

## 3.5 Materials and Methods

### 3.5.1 Experimental data

The experimental data used to validate the EMT model here described were downloaded from the NCBI GEO database [58]. They were obtained measuring in A549 cells, the average level of expression of 29028 genes during EMT induction with TGF $\beta$ . The experiment lasted 72 hours and 3 independent replicates were realized for each of the 9 timepoints considered (Table 3.6).

|                               |                                |
|-------------------------------|--------------------------------|
| Cell line                     | A549                           |
| EMT inducer                   | TGF- $\beta$                   |
| Independent experiments       | 3                              |
| Timepoints                    | 0, 0.5, 1, 2, 4, 8, 16, 24, 72 |
| Tested genes                  | 29028                          |
| Number of probes              | 54662                          |
| Average number of probes/gene | 1.88                           |

Table 3.6: Specifications of the microarray experiment [57] used to validate the EMT model.

The microarray technique used to acquire these data is summarized in Figure 3.3. It consists in using a specifically designed chip, on which a number of short DNA sequences (probes) are immobilized, to detect the presence of the cDNAs of interest, and quantify their concentration. This is realized by using probes whose sequence is complementary to a region of the gene of interest and coupling a fluorescent dye to the sample. A specific reader scans the chip and quantifies the signal. Prior to the hybridization step the sample must undergo a series of procedures aimed to improve its stability and the reliability of the assay. The mRNA is initially purified from the tested cell culture, successively it is retro-transcribed to cDNA to prevent its degradation and finally the cDNA sequences are chemically coupled to a fluorescent dye. The fluorescent signal must be elaborated to compensate for a number of possible distortions, that could be associated to an apparent difference in the recorded signal. Some of them account for variations in the signal due to the different position of the spots on the chip, while others evaluate the noise and the background level. Finally, since the recorded signal is expressed in AU, all the recorded data are normalized with respect to the value of a set of control probes. These pre-elaborated data were organized as an excel spreadsheet and loaded on the NCBI GEO database [58] (accession number GSE17708).

In order to be able to compare these experimental data with the model's results, the expression data relative to the genes in the signature were isolated and a normalization with respect to the initial value was executed. This was obtained dividing each value for a normalizing factor computed as the average of all the data available for that gene, recorded at time 0. Successively all the probes referring to the same marker were condensed in a single value that was considered representative of the effect of EMT on that gene. Specifically, for every timepoint, all the expression data available for each gene were averaged and the standard deviation was computed to evaluate their variability. These values were then compared to the ones obtained with the simulation.

### 3.5.2 *in-silico* data

The simulation of the Markov chain results in a description of the phenotypes evolution during EMT, that allows to infer the key steps of this phenomenon and the influence of the initial population on the simulation results. However a computational model requires a validation, that is a comparative analysis of the *in-silico* results and *in-vitro* data describing the same phenomenon. For this reason the output of the simulations was analysed to determine the fraction of cells, within the population, expressing a defined marker, quantity comparable to the

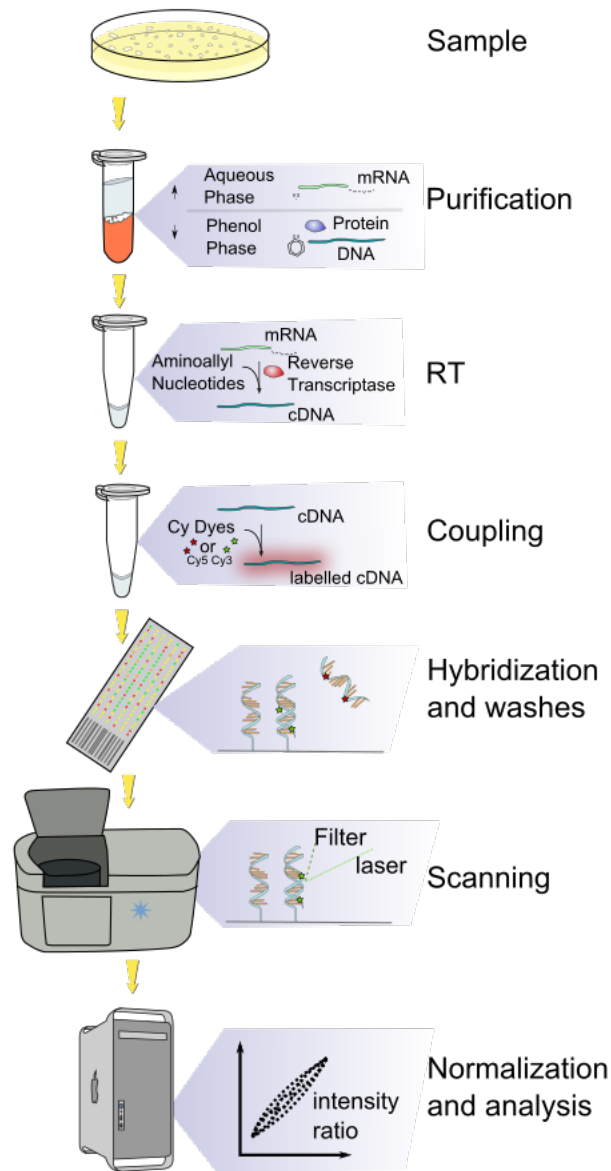


Figure 3.3: Graphical representation of the main steps of a microarray experiment. The mRNA is purified and retrotranscribed, to improve its stability. The resulting cDNA is coupled to a fluorescent tag and then exposed to a specifically designed chip on which DNA probes, complementary to the sequences of interest, are immobilized. The hybridization between the two complementary sequences will make it possible to quantify the concentration of the mRNA of interest, that will be proportional to the recorded fluorescence intensity. To ensure the accuracy of this assay, a set of control probes must be included and an appropriate elaboration must be executed.

average mRNA level determined experimentally through the microarray technique.

Specifically, for every time point, the number of cells expressing each gene of the signature was determined summing the prevalence of all the phenotypes in which that marker is set to 1 and then multiplying for the cardinality of the population ( $5 \cdot 10^5$ ). As for the experimental data all the results obtained with this analysis were normalized with respect to the initial value, leading to a comparison between the variations from the starting condition.

### 3.5.3 Validation

#### Analysis of the Expected Behaviour

This initial step of the validation of the computational model was performed comparing the sign of the variation, with respect to the initial condition, to the expected behaviour obtained researching the literature.

Specifically the number of cells expressing each marker at the end of the simulation was subtracted to the value set as initial condition and the sign of the variation was recorded. The same analysis was repeated for the experimental data, selecting the average level of expression for each marker at a specific time point.

The expected behaviour of each marker during the EMT was determined researching the literature and analysing the signal transduction pathways used to build the boolean network.

#### Study of the Variation of each Marker

After the determination of the coherence between the directions of variation, the corresponding amplitude was considered. Specifically, for each considered dataset, the differences with respect to the initial condition were ranked in decreasing order and the ranks of each marker were compared.

The L1 norm was used to compute the distance between the variations of the same gene measured *in-silico* and *in-vitro* (Eq. 3.1).

$$L1^g = |r_{vitro}^g - r_{silico}^g| \quad (3.1)$$

This value characterizes the differences between the rankings independently of their sign and is thus unable to identify the presence of any bias that might result in consistently higher values for one method. Analysing the results in Tables 3.2 and 3.5, however, the presence of a definite trend is not evident.

The similarity between the rankings was determined counting the number of genes with a distance lower than the average (computed at steady state) and analysing the variations of the average score with the time point.

When the best time point was considered, the coherence between the rankings obtained with the two datasets was also represented as a scatter plot in which a linear fitting, computed through the least square method and using bisquare weights, was used to highlight the correlation between the *in-silico* and *in-vitro* results.

### Identification of the Best Time Point

This analysis was essential to identify which time point was associated with the best correspondence with the results obtained simulating the computational model for 100 iterations.

It consisted in the computation of the correlation between the variation of each marker with respect to the initial condition, obtained with either method. Specifically the difference between the number of cells expressing each gene at the end of the simulation and the corresponding initial value, was computed and correlated to the corresponding values obtained for every time point of the *in-vitro* experiment.

The highest correlation coefficient ( $r=0.75$ ) was determined to be associated with the best conversion factor between the measurement units used in the two experiments to track time.

## Chapter 4

# Quantification of Protein Markers in Single cells using Optical Microscopy

This chapter describes two experimental protocols that were developed to quantify the fluorescent signal emitted by single cells from images acquired with an optical microscope. These instruments, publicly available, can be used to evaluate the concentration of proteins of interest, relying only on general purpose instrumentation that is available in most laboratories.

While serving the same purpose, these softwares differ in their targets. One of them was developed to analyse the behaviour of synthetic gene circuits transformed in *E. coli* cells, through the quantification of one of the most common fluorescent reporters, the green fluorescent protein (GFP), while the other was developed to analyse images acquired during immunofluorescence assays. These experiments are generally executed on human cancer cells, and use specific antibodies to target proteins of interest and quantify their concentration.

### 4.1 Fluorescence Quantification in Single Bacterial Cells

#### 4.1.1 Background

The functionality of synthetic gene circuits is often verified using fluorescent reporters [73, 74], as they allow for a precise and non-invasive quantification of the molecules of interest.

The general workflow of these experiments involves associating a fluorescent tag to the studied mRNA or protein and then illuminating the

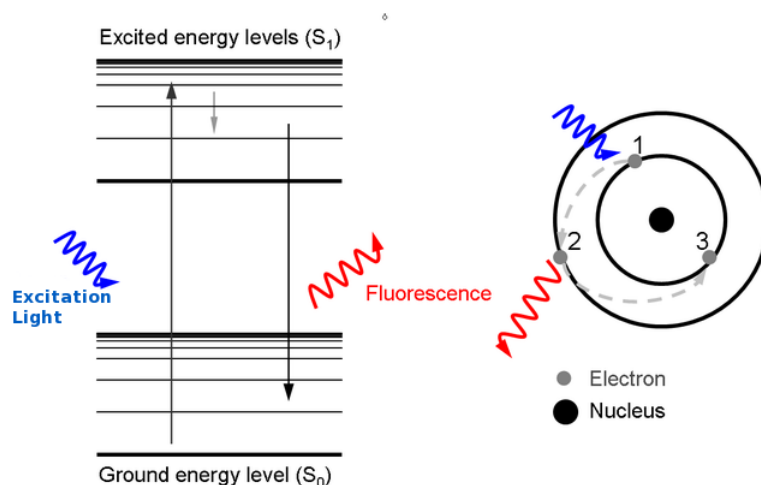


Figure 4.1: Schematic representation of the physical phenomenon that underlies the emission of a fluorescent signal. The absorption of light at a specific wavelength allows the fluorophore to reach an excited state. The excess energy is then released producing a luminous signal at a longer wavelength. Picture reproduced with modifications from [75].

live cell culture with light at a specific wavelength. This stimulation induces in the fluorophore the release of another luminous signal, at a longer wavelength (Figure 4.1), that will be proportional to the concentration of the fluorescent tag and thus to that of the molecule of interest.

A range of photon counting instruments can be employed to record the emitted signal; in synthetic biology the most common are the fluorimetric multiple wavelength plate reader and the flow cytometer. The former is a completely enclosed system that both allows for high throughput results and dynamic measurements of the fluorescence emitted by the same population of cells over time. The latter, on the other hand, features a complex fluidic system that makes it possible to register the fluorescent signal emitted by single cells, but, once measured, the samples are discarded, thus precluding the possibility of executing dynamic experiments on the same cells.

In recent year, flow cytometers have become more and more prevalent, due to the growing body of evidence showing that gene expression noise, i.e. phenotypic variability within an isogenic cell population, is a fundamental component of every biological process [1, 76] granting them robustness in changing environments [77].

However the impossibility, with the flow cytometer, of recording the fluorescent signal emitted by the same cells over time, has led to the devel-



opment of microscopy set-ups that can address this limitation through the use of microfluidic devices and incubation chambers [78, 79, 80].

This approach has the potential of greatly expanding the study of biological noise and its effects on cellular processes, since fluorescent microscopes are general purpose instruments that are available in most laboratories. However to reliably quantify the fluorescent signal these set-ups require a careful and precise calibration, and they also need to be coupled with specific software able to extract and elaborate the signal from the images acquired during the experiment.

In the following will be described an experimental protocol and a software library coded in Matlab [81] and Python [82] that allow to quantify the fluorescence emitted by a bacterial population at single-cell level.

The results obtained applying this method to the optical microscopy set-up of the ICM Laboratory in Cesena, were used to validate the protocol, though a comparison with data acquired with a plate reader and a flow cytometer [83, 84].

#### 4.1.2 Set-up Calibration

As previously mentioned, the correct and reliable quantification of the fluorescent signal emitted with an optical microscope, relies on a precise calibration of the set-up and on the use of specific software to compensate the major distortions that affect the recorded signal.

In the following the three major artifacts that affect microscopy set-ups, i) vignetting, ii) photobleaching, and iii) non-linearities in signal digitalization, together with other iv) minor distortions, will be described and the strategies exploited to compensate them will be detailed.

##### Vignetting

Vignetting consists in a reduced brightness at the image edges caused by imperfections in the lenses system (Figure 4.2). The correct compensation of this phenomenon is important to avoid underestimating the signal emitted from cells that are at the edges of the image.

This aberration can be corrected through the pixel-wise addition of each image to a vignetting one acquired with the same set-up. A vignetting image is created acquiring a picture of a uniformly emitting field and inverting any recorded intensity variation (Figure 4.3). This strategy is simple and widely applicable, since it does not make any assumption on the distortion introduced by the specific set-up. Furthermore the vignetting image doesn't need to be acquired before every experiment, since modifications in the lenses system that could affect this distortion are rare.



Figure 4.2: Example of image affected by vignetting (left panel) and result of the correction of this distortion (right panel). Picture courtesy of Andrea Samoré.

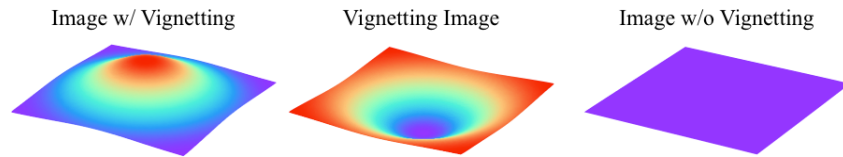


Figure 4.3: Exemplification of the standard procedure for vignetting correction. A vignetting image is obtained complementing an image of an uniformly emitting field acquired with the same set-up used for the experiments. Any intensity variation in this image will be the opposite (but with the same intensity) of those caused by the vignetting. Thus the pixel-wise addition of the vignetting image to any picture acquired with the same set-up will correct any vignetting distortion.

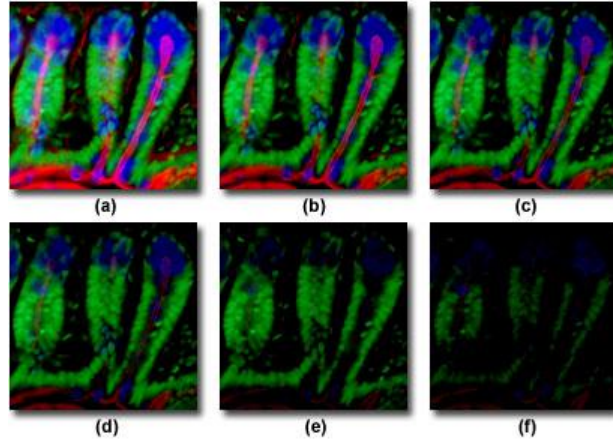


Figure 4.4: Representation of the photobleaching effect. Six different images of the same subject are acquired, with 2 minutes intervals, highlighting the signal's decay over time and the photobleaching's dependency on the fluorophore (the three dyes fade with different dynamics). Image reproduced from [85].

In the presented protocol [84] the vignetting correction is implemented as part of the images pre-elaboration and it expects the vignetting image to be provided by the user as a text file containing the corresponding pixel intensities.

### Photobleaching

Photobleaching is an aberration specific to fluorescence that can be defined as the dimming of the signal over time, due to the photochemical destruction of the fluorophore by the excitation light (Figure 4.4). The significance of this phenomenon is proportional to the exposure time, thus photobleaching compensation is an important step in the analysis of the fluorescent signal acquired with a microscopy set-up. However it is generally omitted in the analysis of flow cytometry data since, with this set-up, every cells is exposed to the excitation light for a very short time. The standard approach for correcting the photobleaching's effect involves measuring the fluorescence decay over time and then fitting the experimental points with an exponential function. This method is described in [86] and allows to identify a function that compensates the photobleaching effect corresponding to a defined exposure time (Figure 4.5).

This method is conceptually straightforward, however timelapse experiments executed with the microscope of the ICM lab highlighted a

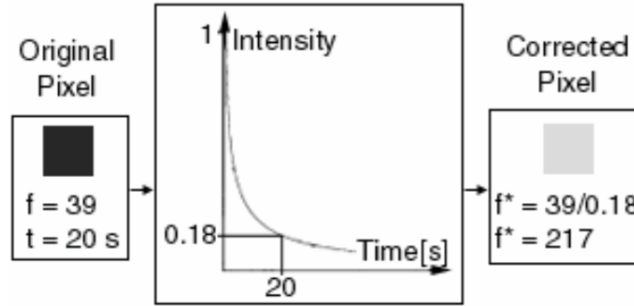


Figure 4.5: Schematic representation of the photobleaching correction described in [86]. The decay of the fluorescence intensity over time was fitted with an exponential curve, that is then used to determine the appropriate correction factor, corresponding to a given exposure time. Picture reproduced from [86].

dependency of the exponential function on a number of parameters, like exposure time, and growth phase, that made the complete characterization of photobleaching demanding and error prone. Furthermore, as shown in Figure 4.4, each fluorophore decays differently, making it necessary to identify a different function for each one employed.

For these reasons the presented protocol [84] uses a different approach to compensate photobleaching. This empirical method consists in setting the maximum number of images that can be acquired from a single slide within a suitable time limit under the continuous exposure to the excitation light. These thresholds can be determined comparing the average fluorescence intensities of the first and last third of images acquired from the same slide and determining which values grant a negligible signal's decay and allow the acquisition of a meaningful amount of data in a reasonable time.

For the set-up used to acquire the data here presented these threshold were set to 15 image/slide and 2 minutes.

### Non-linearity in signal digitalization

The camera response function (CRF) describes the transformation of the scene radiance in the raw images' gray levels. Non-linearities in this analytical relation introduce a distortion in the recorded signal, thereby leading to an erroneous reconstruction of its empirical distribution. The standard calibration technique for the CRF is the radiometric self-calibration method [87, 88] that consists in fitting with a polynomial function the variation of intensity assessed in multiple im-

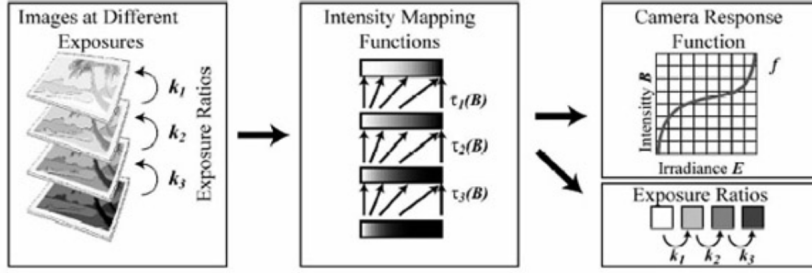


Figure 4.6: Schematic representation of the radiometric self-calibration. Acquiring multiple images of the same scene at different exposure times allows to identify the Intensity Mapping Functions. These relations describe the variation in pixels' intensity between consecutive images and, compared to the ratios between the corresponding exposure times, lead to the identification of the CRF. Image reproduced from [89].

ages of the same field, acquired at different exposure times. Since the radiance of the scene is constant, this relation describes how the luminous signal is modified by the camera as a function of the exposure time (Figure 4.6). Since the CRF depends only on the camera and not on the recorded sample, this aspect of the protocol was developed using bright field images of fixed eukaryotic cells, acquired at four exposure times ranging from 1 to 4 ms. The analysis of these images led to the identification of a third degree polynomial that was inverted with the Cardano's method. During the parameters identification process, the Tikhonov regularization method, as implemented in the Matlab toolbox regtools [90], was used with a threshold automatically selected using the L-plot.

Figure 4.7 summarizes this process, in **a.** is reported one of the sets of images used to determine the intensity mapping functions. These functions compare the gray level variation in the pixels of two consecutive images and the corresponding change in exposure time. Since the radiance of the scene does not change, any difference in these quantities can be attributed to the CRF effect and contribute to its identification. In **b.** the polynomial function that represents the CRF of the camera mounted in the microscopy set-up of the ICM lab is reported (Equation (4.1), where  $IB$  is the image brightness and  $SR$  the scene radiance and in blue in Figure 4.7) together with the experimental data that were used to identify it.

$$IB = -0.066SR^3 + 0.146SR^2 + 0.915SR + 0.005 \quad (4.1)$$

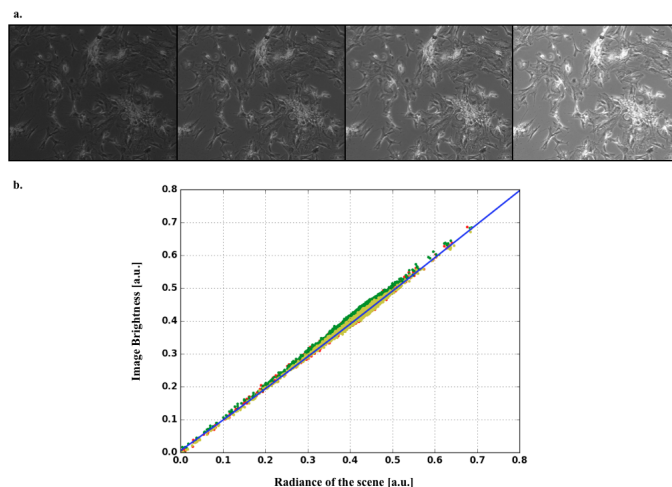


Figure 4.7: Summary of a radiometric self calibration experiment. **a.** Set of images acquired for the CRF calibration. The imaged cells are the same, but the exposure time increases (from 1 ms to 4 ms) from left to right. **b.** Comparagram representing how modifying the exposure time changes the pixels' intensities (dots). The result of the experiment, Equation 4.1, is also reported (blue line).

Since this function can be assumed to be constant, for a given set-up, this characterization needs only to be completed once and the result can be applied to any experiment executed with the same hardware. In the presented method [84] the CRF compensation is carried out in the pre-elaboration phase and was implemented to be fast and efficient. Instead of using the CRF equation, the user is required to provide a text file containing the radiances corresponding to each possible gray level of the image. This allows the program to automatically substitute to each pixel's intensity the corresponding corrected value, without having to compute it every time.

### Minor Distortions

Images acquired with a fluorescence microscope are affected by other minor distortions (e.g. progressive dimming of the arc lamp, temperature fluctuations). As these factors are very difficult to isolate and characterize, and might unpredictably interact, the use of an internal calibrator sample is part of the presented protocol [84]. Therefore, a sample is adopted as a reference throughout all the experiments. Since the calibrator undergoes the same distortions of any other target sample, it provides an implicit correction factor. In addition, the use of a

calibrator allows the meaningful comparison of experiments performed in different days, data acquired with alternative instruments, or samples with divergent fluorescence intensities. In the following maximally induced samples were used as a calibrators, being the ones with the most intense signal.

### 4.1.3 Set-up Validation

#### Description of the protocol

The quantification of the fluorescent signal emitted by single bacterial cells is obtained through the segmentation of the images acquired with an optical microscope and the evaluation of the average gray level of the pixels belonging to each cell. Figure 4.8 summarizes the major steps of the presented protocol; the raw images undergo an initial pre-elaboration that compensates the distortions of the signal, that were described in the previous section. Successively a segmentation routine automatically separates the foreground (the cells) from the background. The heart of this procedure is the zero crossing edge detection algorithm, which is based on the estimate of the null points of the second derivative of the image [91].

To eliminate the spurious edges and identify the ones that more likely represent the outline of a cell, the zero-crossing algorithm is preceded by a smoothing of the image with a Gaussian filter. Once the boundaries of the bacteria have been identified, the algorithm applies a hole filling procedure and then the segmentation is completed by a morphological erosion of the resulting image (Figures 4.8, 4.9). Subsequently, the fluorescence intensity of each cell is computed by averaging the value of all the pixels belonging to a certain cell (identified with a labeling routine on the segmented image).

At the end of the elaboration the output files are saved. A pdf file containing the images at different stages of elaboration and two text files respectively comprising i) the fluorescence intensity of each cell and ii) the density of bacterial cells in each image. To validate the presented protocol [84] a series of experiments were executed, aiming to compare the results of this novel technique to those obtained with the gold standard methods currently employed to evaluate the fluorescence intensity at single-cell level (flow cytometry assay [83]) and the culture's density (optical density measurements).

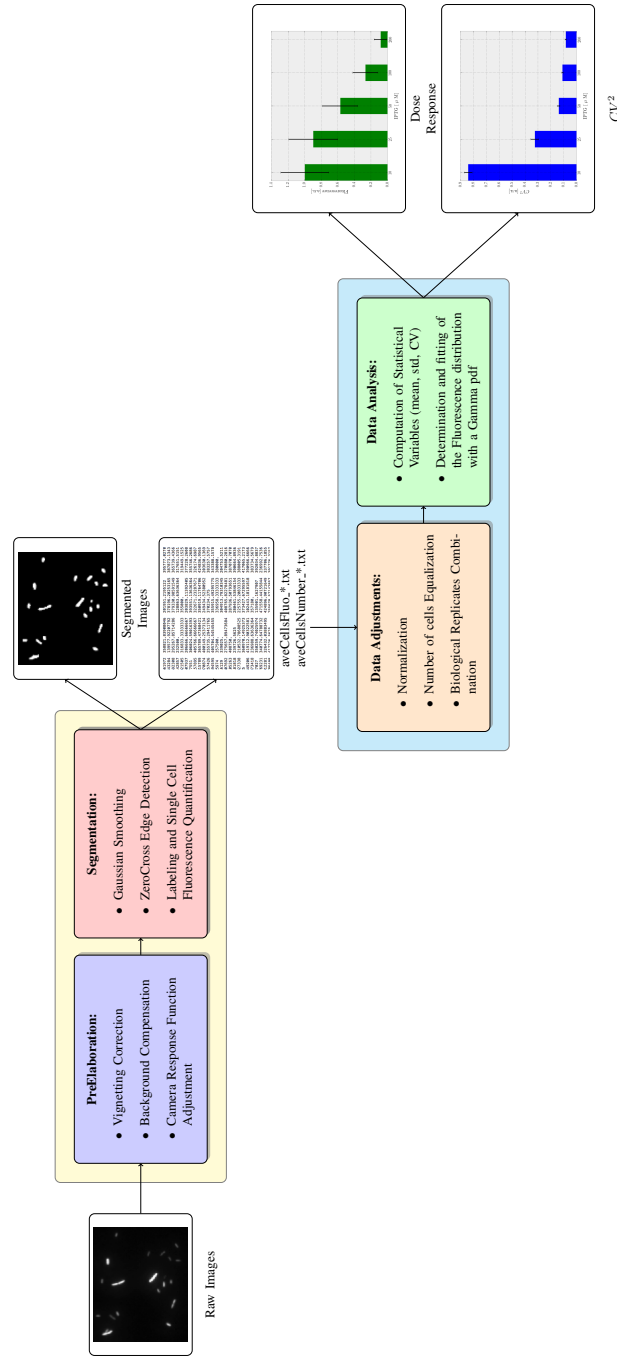


Figure 4.8: Flowchart representing the major steps of the presented protocol [84]. The yellow box identifies the first segment of the analysis in which the images are segmented and the fluorescent signal is quantified. The blue box, on the other hand, describes the steps of the protocol that analyse the fluorescence intensity data and lead to the production of the output graphs.



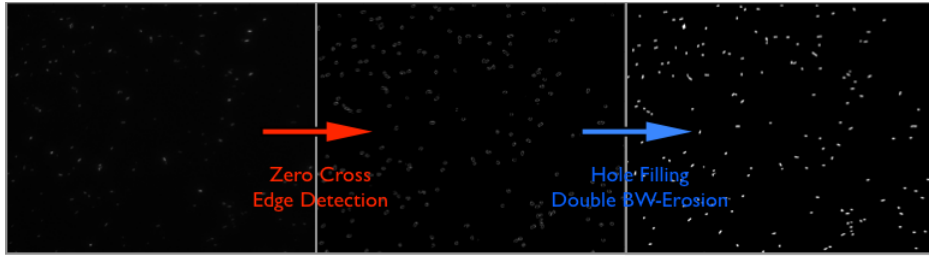


Figure 4.9: Representation of the major steps of the segmentation process. The leftmost image is the one that is acquired during the experiment. The image in the middle is the result of the zero-crossing algorithm [91], where only the edges of the cells are identified as foreground. Finally a hole filling procedure and a double BW-Erosion complete the segmentation.

### Fluorescence intensity quantification at single-cell level.

The single-cell fluorescence intensities were measured in engineered *E. coli* cells where the signal can be transcriptionally induced via exogenous Isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG) (Figure 4.10).

This choice is particularly convenient since a modification of the inducer concentration produces a change in the statistical moments of the distribution, allowing the set-up validation over a wide range of signal's intensities using a single gene circuit, thus avoiding biases introduced by different topologies or environmental conditions. In the following, data will be expressed as average value  $\pm$  standard error (SE). The squared coefficient of variation ( $CV^2$ ) was used to quantify biological noise, since it is a measure of the signal's dispersion around its average value. Both the datasets were normalized with respect to the average fluorescence intensity of the tested circuit at the highest level of induction. The same number of cells ( $\sim 12 \cdot 10^3$ ) was used for the acquisition of each induced fluorescence level with the microscopy set-up, in order to facilitate the comparison among different experimental conditions without distortions introduced by the different cardinality of the populations. This number of cells was above the minimum value required to obtain a stable relation with flow cytometry measurements (Figure 4.11). This graph was obtained dividing the total number of cells acquired with the microscopy set-up in smaller groups of equal cardinality and computing the Pearson's correlation coefficient between the average fluorescence and the  $CV^2$  of these sub-populations and those computed on the complete dataset obtained with the flow cytometer. Populations of just few hundreds cells are able to correctly and reliably capture both the average fluorescent signal and its dispersion around

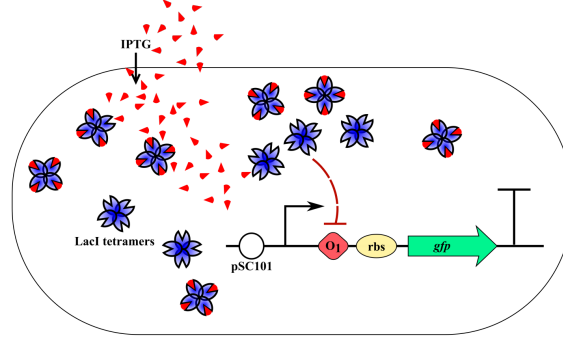


Figure 4.10: Schematic representation of the synthetic gene circuit exploited to compare the single-cell level fluorescence intensities obtained with the presented protocol to those evaluated with a flow cytometer [83]. It exerts a simple transcriptional control through the operator  $O_1$  that, when bound to a LacI tetramer, prevents the transcription of the reporter gene (GFP). The *E. coli* strain used during these experiments (TOP10F') naturally over-expresses LacI, thus the gene circuit is normally switched off. The addition of IPTG to the culture media removes this block in a dose response manner by binding to LacI tetramers and preventing them from operating their repressive function. This Figure was kindly realized by Lucia Bandiera PhD.

the mean.

The core of this method's validation consisted in reproducing the dose-response curve and the relation between  $CV^2$  and induction level obtained with a flow cytometer and presented in [83].

Figure 4.12 a. shows a dose response curve as obtained with both the flow cytometer (blue dots) and the microscope (red triangles). Similar results are provided by the two approaches, with almost superimposed experimental values. The application of the Mann-Whitney u test confirmed that all the induction levels were distinguishable with statistical significance ( $p < 0.01$ ) both with the flow cytometer and the microscopy setup. Figure 4.12 b. highlights the monotonic second degree relation (cyan line,  $MSE = 7.6 \cdot 10^{-5}$ ) between the fluorescence intensities obtained with the two instruments. This relation is almost linear (green line) for values greater than 0.3 a.u. ( $R^2 > 0.99$ ) and shows comparable SEs. Nevertheless, at the lowest induction level this function shows a heavy non-linearity, likely caused by the higher sensitivity of the flow cytometer.

The  $CV^2$  is shown in Figure 4.13 a., where its dependency upon inducer concentration can be observed. The microscopy set-up (red triangles) is able to capture the behavior of this variable of interest, with statistically significant differences in the values associated with distinct

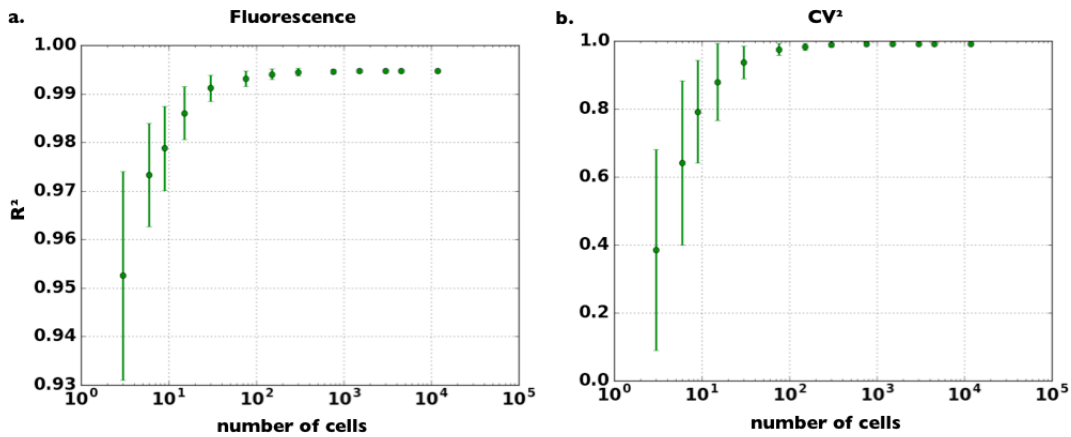


Figure 4.11: These plots analyse how the agreement between the signal registered with a flow cytometer and the one extracted with the proposed protocol [84] changes with the number of cells acquired with the microscopy set-up. In **a.** the average fluorescence is considered, the correlation is very good even for very small populations, however  $R^2$  can vary significantly depending on the specific cells used for the comparison. Increasing the number of cells considered the Pearson's correlation coefficient increases and its variability rapidly decreases. In **b.** the squared coefficient of variation is evaluated. Here this parameter is used to evaluate the dispersion of the fluorescent signal around its mean, and consequently the number of cells used to evaluate it significantly affect the result. However in both cases a few hundreds cells are sufficient to capture both the average fluorescence emitted by the population and its dispersion.

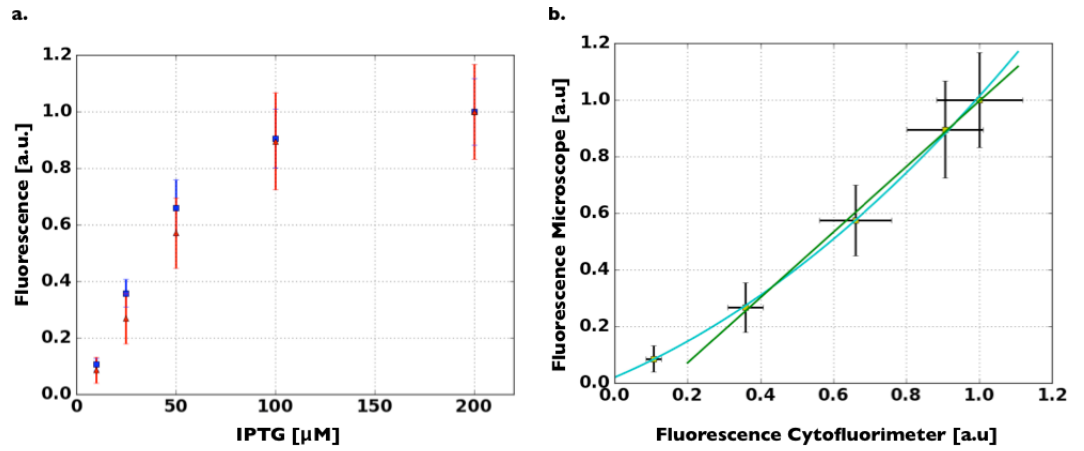


Figure 4.12: Comparison between the fluorescence intensities recorded with the two instruments. In **a.** the average signal, recorded for every induction level and normalized to the value of the maximally induced sample, is reported. The red triangles identify the results obtained with the microscopy set-up, while the blue squares the ones of the flow cytometer. The agreement between the two instruments is very good with comparable average values and SEs. **b.** Correlation plot between the two fluorescence measurements. The correlation is very good  $R^2 > 0.99$ , and the relation between the two quantities is correctly represented by a line on most of the dynamic range (green line). It shows a non-linear behaviour only for the dimmest signal considered, due to the higher sensitivity of the flow cytometer (cyan line).

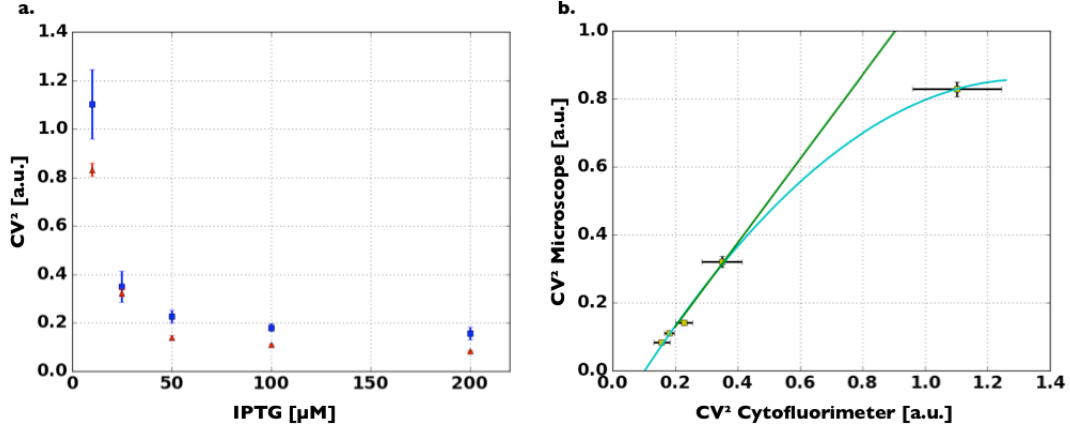


Figure 4.13: Evaluation of the population's dispersion around the mean. The  $CV^2 = (\frac{\sigma}{\mu})^2$  is used to evaluate this aspect. In **a.** the  $CV^2$  computed for every induction level with both the instruments tested is reported. The presented protocol [84] is able to reproduce the decreasing trend of this variable, with the highest difference corresponding to the lowest induction level. This is also clear in **b.** where the correlation between the two measures is explored. The  $CV^2$ s are linearly related on most of the tested dynamic range (green line), only the highest value, corresponding to 10 mM of IPTG, shows a heavy non-linearity (cyan line).

induction levels of the tested synthetic gene circuit (Mann Whitney u test,  $p < 0.01$ ). These results are coherent with the measurements performed using flow cytometry (blue dots in Figure 4.13 a.), as highlighted by the correlation graph in Figure 4.13 b. As already remarked for the dose-response curve (Figure 4.12 b.), the second-degree relation between  $CV^2$  extracted with the two approaches is linear on most of the dynamic range ( $R^2 > 0.99$ ) and shows comparable SEs. The lowest induction level, corresponding to the highest  $CV^2$ , is responsible for the nonlinear behaviour ( $MSE = 1.8 \times 10^{-4}$ ) that is associated to a higher dynamic range of the flow cytometer, which is able to better capture dimmer fluorescent signals.

### Identification of the culture's density

A complete characterization of a gene circuit requires, beside evaluating the intensity of the fluorescent signal emitted by individual cells in a population, an estimate of the culture's density. This is fundamental since it has been demonstrated that growth phase and nutrient abundance influence significantly gene expression [92, 93].

Thus an important functionality of the presented method [84] is the

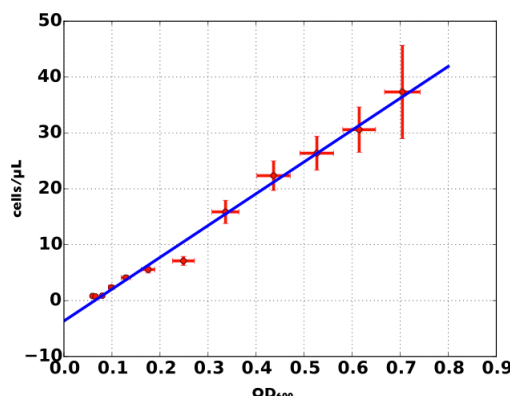


Figure 4.14: Comparison between the culture density estimation obtained with the presented protocol and the corresponding  $OD_{600}$  measurements. The relation between these variables is linear and  $R^2=0.996$ . Figure reproduced from [94].

ability of quantifying the population's density using the same images from which the fluorescent signal is extracted. It simply counts the number of segmented cells in each image and relates it to the volume used to prepare the slide. Correcting the resulting density if the culture was concentrated/diluted to obtain the optimal image filling, that is maximize the number of cells in the image while reducing the number of superimposed bacteria.

This measure shows a very good agreement with the optical density ( $OD_{600}$ ) measured using a plate reader, that can be considered the gold standard measure of cell culture density (Figure 4.14). To compare these two measurements the growth curve of the same cell culture, transformed with the circuit in Figure 4.15, was measured over a period of 4 hours both with the plate readers (triplicate measures for each time-point) and the microscopy set-up (an average of 16 image/-time point). These cells constitutively express the fluorescent reporter allowing for the correct quantification of the cell density even at the beginning of the growth curve. The correlation graph in Figure 4.14 clearly shows the linear relation between the cell density evaluated from the images and the  $OD_{600}$  measurements ( $R^2=0.996$ ).

## 4.2 Fluorescence Quantification in Eukaryotic Cells

### 4.2.1 Background

The evaluation of gene expression in eukaryotic organisms, especially human cells, has traditionally focused on population data and the mea-

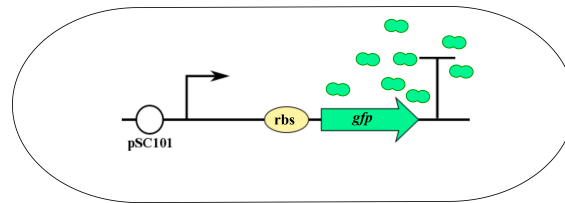


Figure 4.15: Schematic representation of the gene circuit used to compare the OD measurements, obtained with a plate reader, and the cell density estimated from the microscopy images. It expresses GFP constitutively, thus producing a constant fluorescent signal that allows the correct identification of the growth curve.

sure of the average level of mRNA or protein concentrations in a large number of cells. This approach, while capable of measuring variations in gene expression that affect a large part of the population, doesn't capture more subtle changes that act on a small fraction of the cells and have been associated to a number of biological processes, like cancer progression [95, 96, 97, 98], stem cells maintenance [99], embryogenesis [100] and aging [101].

In recent years experimental techniques able to quantify the concentration of specific mRNAs or proteins at single-cell level have been developed [99, 102, 103, 104]. Most of these techniques focus on mRNA quantification and build on the recently developed next generation sequencing that ensures high throughput and overcomes the limitations of RT-PCR methods, that require a rigorous set of controls and have significant limitations on the number of genes and cells that can be tested in the same experiment [105].

These techniques, however, are significantly less widespread than more traditional approaches like RT-PCR and this limits the study of inter-cellular variability in eukaryotic cells, and especially in human cells.

Building on the experience acquired developing a protocol for the quantification of fluorescent signal emitted by single bacterial cells, a method for the analysis of images acquired during immunofluorescence assays, able to quantify the concentration of specific proteins in human cancer cells, was developed. Immunofluorescence assays are a classical microbiology technique that exploits antibodies conjugated with fluorescent dyes to bind specific proteins of interest, thus leading to a direct proportionality between the protein's concentration and the fluorescent signal.

Unlike the assay described in the previous section, fixed cells are imaged during immunofluorescence assays, thus preventing the measurement over time of the signal emitted by the same cells. Another difference is

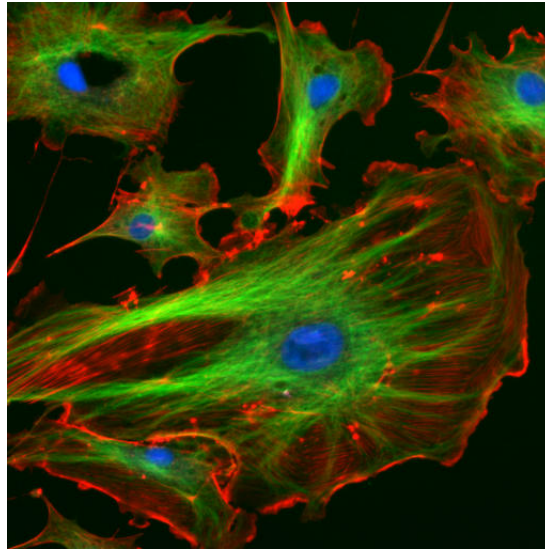


Figure 4.16: Example of an image resulting from an immunofluorescence assay. The blue dye DAPI is used to mark the cells' nuclei while the green and red signals are correlated to the concentration of the proteins of interest. Beside quantification of the targeted molecule, this assay can also determine its localization within the cell, information tightly connected to the protein's function. Image reproduced from [106].

the use of an additional dye, beside the fluorescent antibody, that can bind the DNA and identify the cells' nuclei. This additional information allows to easily count the cells in each image and it is fundamental for the presented segmentation algorithm.

In Figure 4.16 is reported an example of the images resulting from immunofluorescence assays. The cells' nuclei are marked in blue (4',6-Diamidino-2-Phenylindole, Dihydrochloride -DAPI- dye) while the protein of interest, in this case E-Cadherin (CDH1), is shown in green. Multiple antibodies, combined with different fluorescent dyes, can be used at the same time to measure the level of expression of different proteins in the same cells, only care must be taken in the selection of the fluorophores, to avoid crosstalk between different signals. Another information that can be extracted from these images is the proteins' location within the cell. Unlike bacterial cells, where the fluorescent signal was distributed uniformly, proteins in human cells have specific collocations that are tightly connected to their function. In the following different approaches will be applied to quantify the signals from the nuclear or cytoplasmic/membranous compartments, thus allowing to differentiate the concentrations of the same protein in different compartments.



### 4.2.2 Algorithm description

In immunofluorescence assays each signal is generally acquired separately and then combined to obtain images like the one in Figure 4.16. This is exploited by the presented protocol that, analysing them separately, can apply different procedures to images with different characteristics.

Specifically the DAPI signal, or fluorescent dyes which signal concentrates in the nucleus, can be easily identified with a procedure similar to the one implemented in the previously described bacterial algorithm [84], in which the cells are identified with an edge detection algorithm complemented with some morphological operations. On the other hand, signals that are mainly localized in the cytoplasm or on the membrane, are better localized by techniques that do not rely on the presence of a stark contrast between the foreground and the background. One of such techniques is the Watershed transform [107] that can be visualized as placing water sources in specific regions of the image (markers in Figure 4.17) and then building watersheds (drawing edges) where water from two different sources meets. As clearly shown in Figure 4.17 the number and positioning of the markers is fundamental for achieving the correct result, using only two markers (Figure 4.17 b., leads to the fusion of the two rightmost basins, that are correctly segmented in Figure 4.17 a.

Markers positioning in the presented algorithm is achieved by placing one marker over each cell's nucleus identified segmenting the images acquired over the DAPI channel. As previously mentioned the nuclei of the cells can be identified with an edge detection algorithm, due to the high contrast between the foreground and the background. Thus the markers are identified with a modified version of the bacterial algorithm described in [84], in which the parameters have been modified empirically, to compensate for the difference in dimension between the bacterial cells and the eukaryotic cells nuclei. This information is then used to segment the cytoplasm of the cells using the watershed transform.

Once the cells have been segmented, the signal emitted by each cell is quantified as the average intensity of the signal in the region identified with the watershed transform, if the protein is localized in the cytoplasm or on the membrane. For nuclear proteins the signal is extracted from the area segmented with the modified version of the bacterial algorithm. As detailed in the previous section, the correct quantification of fluorescent signals with an optical microscope is dependent on the compensation of the distortions introduced by the acquisition system. All the calibration strategies previously described are also implemented

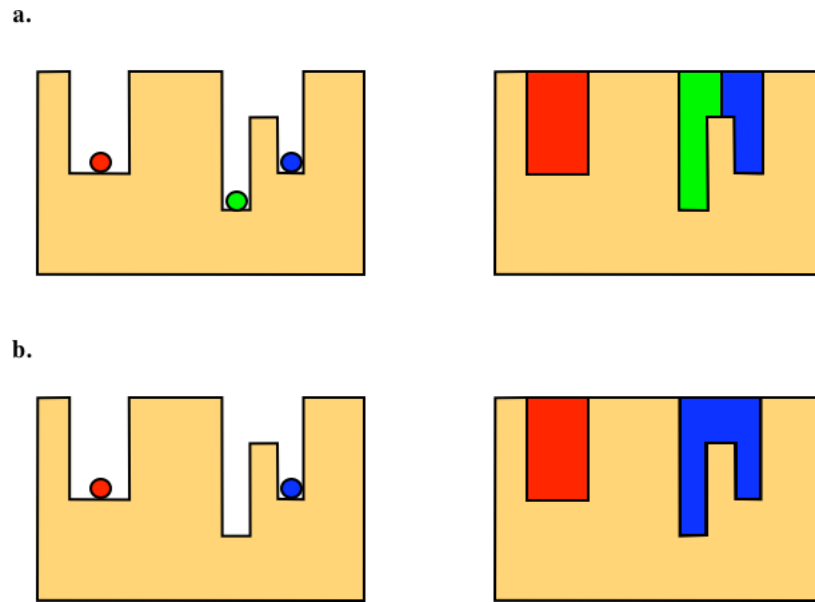


Figure 4.17: Graphical representation of the Watershed transform, each foreground object is identified with a marker, that acts as a water source that fills the neighbouring region. The algorithm builds watersheds (edges) on the lines where water from adjacent basins connects. In **a.** the positioning of three markers leads to the correct identification of all the minima, while in **b.** only the leftmost region is correctly segmented due to an incorrect positioning of the markers.

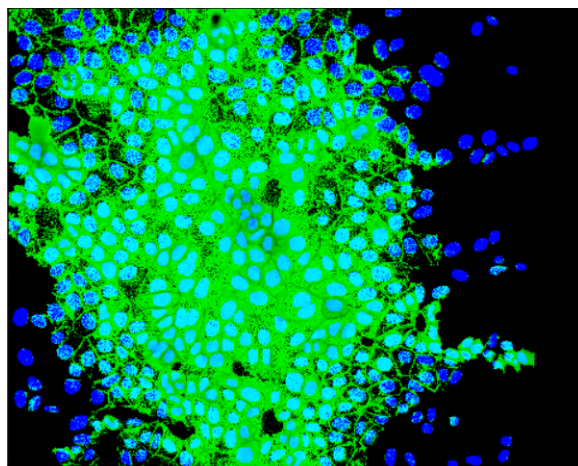


Figure 4.18: Image in Figure 4.16 segmented with the presented method. The software automatically combines the segmented images acquired over the different channels and produces the output of a classic immunofluorescence assay.

in this protocol, thus ensuring the accurate detection of the signal and full compatibility between the two methods.

### 4.2.3 Results

The protocol for the quantification of fluorescent signals in eukaryotic cells has been used to test the changes in the expression of CDH1 induced in MCF7 cells by  $TGF\beta$ , a potent inducer of Epithelial to Mesenchymal transition (EMT) *in-vitro*. MCF7 are human breast adenocarcinoma cells, isolated in 1970 from the breast tissue of a 69 year old Caucasian woman, that exhibit an epithelial phenotype [108]. Thus the addition of  $TGF\beta$  is expected, over time, to reduce the expression of CDH1, an hallmark of the epithelial phenotype, as the cells undergo EMT.

In Figure 4.18 one of the segmented images obtained with the presented protocol is reported. The software automatically combines the segmented images acquired over the different channels to produce the combined image that is generally considered the output of immunofluorescence assays. Additionally a series of text files is produced, in which the fluorescence intensities emitted by each cell and the corresponding number of cells are reported. These data are organized so that different conditions and different time points are clearly distinguishable.

Figures 4.19 and 4.20 report the results of this experiment. The former shows the change in average cellular density and mean fluores-

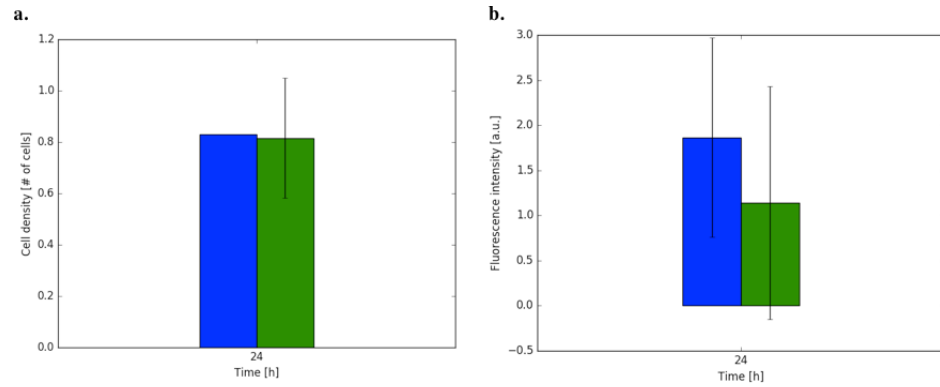


Figure 4.19: Bar graph showing the changes in cellular density and average fluorescent intensity over 24 hours. In blue is represented the control, while green identifies the population treated with  $\text{TGF}\beta$ . Since both the populations derive from the same cell culture only the control condition was tested at  $T=0$  h and it is used as a normalizer in the subsequent analysis. **a.** The cellular density doesn't vary significantly during the experiment, with only a reduced variability at  $T=24$  h and the confirmation that  $\text{TGF}\beta$  doesn't have a significant effect on the growth rate, within the tested conditions. **b.** The fluorescence intensity of the control increases between the two time-points, this is probably caused by the formation of a more complex cell-cell adhesion structure that requires an higher level of CDH1. The treatment with  $\text{TGF}\beta$ , induces EMT in part of the population, thus leading to an average signal comparable to the one recorded at  $T=0$  h, but with higher standard deviation.

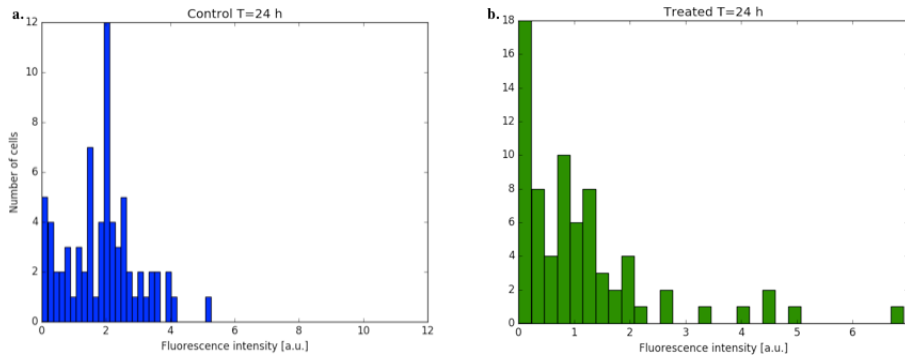


Figure 4.20: Distribution of the fluorescent signal at T=24 h. **a.** In the control condition, CDH1 is more uniformly distributed, with most of the cells emitting a signal about twice as bright as the control at T=0 and a maximum recorded intensity about 5 times higher than the normalizer. **b.** The cells treated with  $\text{TGF}\beta$ , on the other hand, are more diverse, with a CDH1 range of over 10 folds, even if most cells express low levels of CDH1. This result show that EMT, at least in these first phases of its induction, is a process that interests only part of the treated population.

cence intensity after 24 hours of induction with  $\text{TGF}\beta$ . Blue and green represent the control and the treated conditions, respectively and since the cells were seeded from the same population, only the control condition was tested at T=0 h and used as normalizer in the subsequent analysis.

The cellular density doesn't change between T=0 and T=24 h, but its variability decreases, hinting to a possible bias in the selection of the fields imaged at the first time point that led to an overestimation of the cell density. But probably the most important aspect of the left panel of Figure 4.19 is that it shows that  $\text{TGF}\beta$  has no effect on the growth rate of the MCF7 cells at 24 h.

The average fluorescence intensity of the control increases during the experiment as CDH1 is a calcium-dependent cell-cell adhesion protein and the additional 24 hours of culture give the cells enough time to build a denser network, that requires an increased CDH1 expression. Supplementing the media with  $\text{TGF}\beta$  prevents this from happening, thus leading to a CDH1 level similar to the one recorded at T=0 h, but with an higher variability, showing how EMT, at least in the first phases of its induction, is a phenomenon that influence only a sub-population of the cells treated with  $\text{TGF}\beta$  (right panel Figure 4.19). Figure 4.20 analyses more in detail this aspect, showing the distribution of the fluorescent signal at 24 hours in both conditions. In the control, represented in blue, the cells show a more uniform fluorescence

distribution, with a maximum intensity of about six fold higher than the average fluorescence at  $T=0$  and a median fold increase of two. The treated condition, on the other hand, has a maximum recorded signal that is over 10 times the one recorded 24 hours before, but most of the cells emit an almost undetectable signal, remarking how EMT is a phenomenon that, at least in its first phases, occurs in only a sub-population, the one that shows a reduced expression of CDH1.

### 4.3 Discussion

The computational tools presented in this chapter contribute to the development of accurate methods and techniques to extract quantitative information from biological systems. Specifically they aim to quantify the concentration of proteins of interest at single-cell level, thus allowing the evaluation of phenotypic variability within an isogenic population. This aspect is of great importance both for prokaryotes, where it is associated to robustness in changing environments [77] and eukaryotes, being involved in cancer progression [95, 96, 97, 98], stem cells maintenance [99], embryogenesis [100] and aging [101].

Both methods have been developed using image analysis techniques to segment the cells in pictures acquired with an optical microscope and quantify the signal emitted by each element of the population. This approach, while being less efficient than other techniques (e.g. flow cytometry) in terms of throughput and speed of the analysis, is decisively less expensive and doesn't require specific instrumentation. As a consequence the presented tools have the potential to significantly expand the study of cellular noise and phenotypic variability, making it accessible to small laboratories that generally have limited resources.

The algorithm developed for bacterial cells, described in [84], has been completely validated and its performances were determined to be equivalent to those of a flow cytometer on a large dynamic range. This tool, freely available at [109] and released under the GNU public licence (GPL v2), could be used by synthetic biologists to functionally characterize their newly developed circuits or to study naturally occurring phenomena over time, since this method allows for the continuous monitoring of the same bacterial population over time.

The tool developed for human cells, on the other hand, was only tested on a limited number of images. This preliminary validation, however, was able to determine a change in the average recorded intensity and in the signal's distribution. Both these results are coherent with the expected ones, suggesting that this algorithm, once thoroughly tested, could be used to study phenotypic variability in a cancer cell

line population, addressing one of the most important limitation of the techniques most widely used, their inability to distinguish the signal emitted by each cell.

Another important result of this chapter is the definition of a standard protocol for the calibration of optical microscopes. This process, described in detail and adaptable to any set-up, is required for the generation of quantitatively accurate data and can be applied, independently of the segmentation tools here presented, to correct the aberrations introduced by these acquisition systems.

Altogether this chapter demonstrates how general purpose instrumentation, commonly available in most laboratories, can be coupled with *ad-hoc* acquisition protocols and specific image and data analysis tools to produce accurate and quantitative results that have been demonstrated to be equivalent to those of instruments specifically designed to quantify gene expression at single-cell level.

## 4.4 Material and Methods

### 4.4.1 Set-up calibration

As described in the previous section the signal acquired through microscopy set-ups is affected by three main aberrations: vignetting, photobleaching and non-linearities in signal's digitalization. All these factors were considered when developing the presented protocols and correction strategies were integrated within the analysis.

#### Vignetting

This distortion was corrected by subtracting to each image the complement of a vignetting image acquired with the same set-up (Figure 4.3). This picture was obtained imaging a green fluorescent reference slide, that, being saturated with the fluorophore, grants the uniformity of the recorded signal. As mentioned previously, the set-up used to acquire the data analysed with the bacterial protocol was determined to have a negligible vignetting, but the protocol and the corresponding software was developed so as to be fully adaptable to microscopy set-ups requiring the correction for this aberration.

### Photobleaching

Photobleaching is generally compensated through a calibration function, that describes how that particular fluorophore degrades with time as it is exposed to the excitation light (Figure 4.5, [86]). Although pretty straightforward, this strategy was unable, in our set-up, to isolate the photobleaching effect from other confounding factors like growth phase of the culture or exposure time. Furthermore each fluorophore decays with a different characteristic, thus requiring to be characterized independently. To overcome these limitations in the presented protocols this aberration is corrected empirically by setting two thresholds, one on the maximum number of images that can be acquired from the same slide and the other on the longest time of exposure of the same cells to the excitation light. This approach was demonstrated to be associated with a negligible decay of the fluorescent signal. The comparison of the average fluorescence intensity recorded in the first third of images, acquired from the same slide, and that extracted from the last third, lead to the identification of 15 images/slide and 2 minutes as suitable values for the aforementioned thresholds. As shown in the results section these values also allows for the acquisition of a significant number of cells within a reasonable time.

### Non linearities in signal digitalization

This distortion refers to any modification of the signal introduced by the system used to acquire it and can be compensated through the identification of the CRF. This is a function that describes how the radiance of the scene is transformed in the gray levels of the image and is specific of the utilized camera. In the presented protocol this effect is compensated applying the radiometric self calibration [87, 88]. As shown in Figure 4.6 it consists in acquiring different images of the same scene at increasing exposure times, and then comparing the ratio between the gray levels of the same pixel in adjacent images to the corresponding increase in shutter speed. Any difference between these values, is an effect introduced by the camera, since the signal emitted by the scene doesn't change, and it can be used to identify a polynomial function that, once inverted, allows the compensation of the CRF. Since the grade of this polynomial is not known *a-priori*, in [87] it is suggested to repeat the analysis with functions of different grade and then computing an error function to determine the one that best approximates the transformation introduced by the camera. This strategy was implemented, using the same set of images to identify 10 functions, with grade ranging from 1 to 10, and an error function, re-



ported in Equation (4.2), is used to select the one that best reproduces the distortion introduced by the camera.

$$E_g = \sum_i \sum_p \left| \frac{CRF_g(p)^i}{CRF_g(p)^{i+1}} - \frac{T_{exp}^i}{T_{exp}^{i+1}} \right| \quad (4.2)$$

This equation, where  $p$  is an index that varies over all the image's pixels,  $i$  identifies the current image and  $g$  is the grade of CRF, determined that the third degree polynomial was the one with the lowest error, for the set-up used to acquire the results presented in the previous section. For this experiment were used images of fixed eukaryotic cells, since the CRF does not depend on the acquired signal, but only on the utilized hardware. Furthermore care was taken in the selection of the exposure times, to avoid saturation, since in the radiometric self calibration pixels that are saturated or exhibit a non monotonic relation with the exposure time are excluded from the analysis. All these distortions are compensated in the **preProcess** function of the presented softwares that, executed before segmenting the images (Figure 4.8), ensures the correct determination of the signal emitted by each cell. This function also applies a background correction, aimed to increase the uniformity of the cell-free regions, thus reducing the detection of spurious gradients by the segmentation routine. The applied procedure is described in [110] and consists in subtracting to each image its morphological opening, obtained using a structuring element of the same size or bigger than the foreground objects. In the bacterial protocol a square structuring element of size 80 pixels was shown to effectively remove the background in all the tested images. Another advantage of this technique is the lack of hypothesis on the background's effect on the image, thus making it applicable to any microscope set-up.

#### 4.4.2 Fluorescence Quantification in Single Bacterial Cells

##### Inputs and outputs

All the functions that are necessary for the segmentation of fluorescent bacterial images with the presented protocol are collected in the class **microscopeassay** coded in Python 2.7.11 and released under the Gnu Public Licence (GPL v2). The constructor of this class has as only one mandatory argument, that is the absolute path of the folder containing the images to analyse. It has also other three arguments (replicates, imageDims and volumePerSlide), that need to be provided only if they differ from the default values, 2, 1024x1280 and 3  $\mu$ L, respectively. In order for the system to load and properly elaborate the images,

they need to be stored in 8-bit tiff format and coded in the RGB color space. Furthermore, they need to be organized in a folder structure that separates them according to biological replicate, circuit, level of induction and acquisition time point.

In particular, each folder name needs to be encoded by following the scheme  $[d\_c, tP\_iC\_bioR\_expT\_cF]$ , where:

- $d$  is the experiment date,
- $c$  is the identification string of the gene circuit and
- $tP$  identifies the acquisition time point.
- $iC$  represents the inducer concentration,
- $bioR$  identifies the biological replicate,
- $expT$  indicates the shutter speed,
- $cF$  is the factor by which the volume of the culture was reduced before acquisition.

This allows for the automatic identification of the main characteristics of the experiment, the correct association between data and experimental condition, the determination of culture density and the application of the appropriate correction for different exposure times.

The outputs of the segmentation algorithm are three files, a pdf and two txt documents. The former displays the analysed images at different stages of elaboration, while the latter report, respectively, the fluorescent signal emitted by each tested cell, and the number of cells identified in each image. These files are the input of the **microscope-dataanalysis** class, this set of functions is responsible for the analysis of the collected data and the production of the final graphs. As they may vary considerably between different experiments, this class might not comprise all the necessary functionalities, but it has functions that compute the most common statistical moments (average, standard deviation, skewness and kurtosis) and that quantify the noise of the population, through the evaluation of the  $CV$  and the  $CV^2$ . This class is fully compatible with the most common Python graphical library, matplotlib, allowing for a flexible albeit rigorous analysis of the data.

### Segmentation algorithm

As described in the previous sections, the bacterial cells are segmented applying the Zero Crossing edge detection method [91], that places the edges where the second derivative of the gradient of the image is zero. This strategy is exemplified in Figure 4.21 where it is shown how the differential operator affects the intensity profile of an edge in which the foreground is lighter than the background. The intensity profile, in

blue, increases in the transition between the background and the foreground and decreases when shifting from foreground to background. The first derivative would identify these changes in intensity as a maximum and a minimum, respectively, while in the second derivative (in red) they all correspond to the points in which the function crosses the zero. This algorithm is preceded by the filtering of the image with a Gaussian kernel with standard deviation of 2 pixels, this step smooths the image, thus reducing the detection of spurious edges.

The segmentation is completed by a hole filling procedure and a double black and white erosion, that is necessary to equalize the pre and post segmented cell size. Successively, the images are labelled, this procedure assigns a unique identifying number to each segmented region and is instrumental to the quantification of the fluorescent signal, since it allows to easily extract the pixels belonging to each cell, whose average value is one of the outputs of the presented protocol.

### Data analysis

The analysis of the fluorescent data is achieved through the functions of the **microscopedataanalysis** class. This part of the protocol is less structured than the segmentation of the images, to make it adaptable to different set-ups and experiments. However it implements a number of general purpose functions that can be combined to obtain the final results of the experiment.

The constructor of this class requires two arguments: the absolute path of the folder containing the text files produced during the segmentation of the images and a list containing the dates of the experiments that the user wants to analyse. The execution of this function loads the data from the text files and organize them in a format that is recognized by the other functions of the class.

A meaningful comparison between different conditions and/or instruments requires the data to be adjusted for a number of confounding factors, like the population's cardinality and the minor distortions that affect microscopy set-ups and are detailed in the calibration section. For this reason the class **microscopedataanalysis** includes the function **standardAnalysis** that executes three main functions:

- It normalizes the data, with respect to the average intensity recorded for the gene circuit and the condition identified by the user as normalizers.
- It equalizes the cardinality of the population, to the largest possible value given the data or to the one provided by the user.

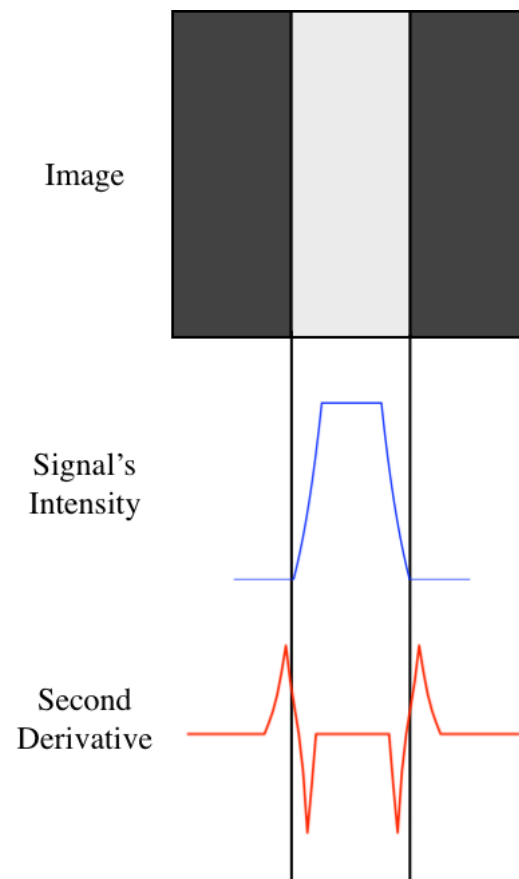


Figure 4.21: Common edge detection techniques apply the differential operator to the image, since edges are characterized by sharp intensity variations. The first row of this figure presents an example of edge in which the foreground has a higher intensity, when compared to the background. This is also exemplified in the intensity profiles, shown in blue. The last row represents the effects on the intensity profiles of the second derivative. The edges of the image are identified as the points in which the function crosses the zero line.

- It combines all the data acquired for the same gene circuit and in the same condition.

All these functions can also be executed independently or combined differently depending on the requirements of the specific experiment. Other two very important functions that are coded in the **microscope-dataanalysis** class are **computeStatistics** and **ComputeCV** that computes some of the most common statistical parameters for the acquired data. The former determines average value, standard deviation, skewness and kurtosis, while the latter calculates CV and  $CV^2$ .

These functions were used to analyse the data presented in the Result section (Figure 4.12, 4.13, 4.11), while the flow cytometry results were extracted from [83].

### Culture density estimation

The determination of the cellular density is an important aspect in the study of biological processes as, together with the growth phase, it affects the system's functionality. The presented protocol estimates the culture density by relating the number of cells segmented in each image, to the volume of culture used to prepare the slide. These values have been demonstrated (Figure 4.14) to be equivalent to the measurement of the optical density. The latter is an absorbance measure that consists in illuminating the sample with light at a specific wavelength, that is absorbed by the cell membrane, and quantifying the radiation that passes through the sample (Figure 4.22). The more cells are in the culture, the less intense will  $I_1$  be, thus making the  $OD_{600}$  value proportional to the cellular density.

To evaluate the equivalence between  $OD_{600}$  and the cellular density extracted from the microscopy images, a comparison experiment was set up. It consisted in following the growth curve of two bacterial *E. coli* populations over a period of 4 hours. The tested cells were transformed with the gene circuit in Figure 4.15, that constitutively expresses GFP and the two populations differed only for the promoter's strength. After diluting the over-night cultures to an  $OD_{600}$  of 0.05 the growth of both populations was monitored every 30 minutes both taking pictures of the cells (an average of 16 images for each time point and condition) and measuring the  $OD_{600}$  (in triplicates) with a plate reader. The images were then segmented and the average cell count, divided by the volume used to prepare the slide, corrected for any dilution/concentration factor applied to the culture, was compared to the corresponding  $OD_{600}$  measurements, showing a very good correlation ( $R^2=0.996$ ). This experiment demonstrated how the presented protocol is able to

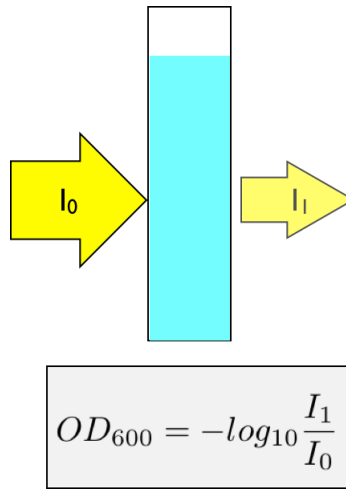


Figure 4.22: Exemplification of the functioning principle of the  $OD_{600}$  measure. The sample is illuminated with light at a wavelength of 600 nm ( $I_0$ ), that is absorbed by the cell membrane. Thus the transmitted light ( $I_1$ ) will be inversely proportional to the number of cells in the sample. This signal is measured and then related to the intensity of the incident light. The logarithmic scale is used to improve the comparison of signals that have large dynamic range.

correctly estimate the density of the bacterial culture, through the evaluation of the average number of cells segmented for each condition, divided by the volume used to prepare the slide.

### Experimental protocol

- Experimental model:** The data presented in this chapter were obtained testing synthetic gene circuits composed of standard biological parts in the BioBrick format. Gene circuits conformed to the Standard Assembly 10, and were transformed in TOP10F' *E. coli* cells. These were cultured in M9 medium complemented with the antibiotic ampicillin and glucose as a major carbon source. The circuit used to quantify single-cell fluorescence level, previously characterized in [83], includes a reporter fluorescent signal (GFP) downstream of an operator site for the lactose repressor, and consequently it can be transcriptionally induced via exogenous IPTG (Figure 4.10). Before measuring the fluorescence levels each culture was diluted to an  $OD_{600} = 0.05$  and grown under orbital shaking for 3 hours, to reach the mid-exponential phase of growth, at  $37^\circ C$  in 5 ml of M9 medium in the presence of the appropriate concentration of IPTG. Cell fluorescence signal was

subsequently measured with: a) a fluorescence microscope and b) a cytofluorimeter. Three biological replicates were considered for each tested condition. To compensate the bias introduced by the time lag between the testing of the first and last sample, the acquisition order was varied among the biological replicates. The sequences were determined so that the sum of the rankings of each sample over the biological replicates was equal. To further limit the deviation from the desired condition, after the beginning of the acquisition the cultures were stored at  $4^{\circ}\text{C}$ .

- **Microscopy set-up:** An inverted Eclipse TE2000-U (Nikon) microscope equipped with a DS-Qi1Mc (Nikon) (Table 4.1) monochrome digital cooled camera was used to collect brightfield or fluorescent images of culture samples through an S-Fluor 40x 0.9 NA oil/water (Nikon) objective. The proprietary Nis Elements Documentation v 4.20 software (Nikon) was used for image acquisition.
- **Image acquisition:** To prepare the sample for image acquisition, 500  $\mu\text{L}$  from each cell liquid culture sample were spun down and resuspended in 100  $\mu\text{L}$  of sterile PBS to reduce the background autofluorescence and to maximize the cardinality of the sampled population while preserving an optimal field of view coverage. A volume of 3  $\mu\text{L}$  of this cell suspension was dispensed over a glass slide and sealed with a coverslip. A minimum of 70 images out of 6 distinct slides were acquired for each sample. During image acquisition, the shutter speed is heuristically defined to distinguish clearly the cells while avoiding loss of information due to saturation. When cells with different average fluorescence are acquired, this parameter needs to be modified in order to correctly capture both the minimal and the maximal fluorescence intensity values, which may hamper the comparison among samples acquired with different exposure times. Having characterized the camera response function, however, permits to express the fluorescence values in terms of normalized irradiance rather than pixels intensities. This allows to reliably compare samples acquired with a different shutter speed simply dividing the normalized irradiance by the exposure time set during the acquisitions, thus restoring the right relationship between different samples.



|                                |  |
|--------------------------------|--|
| Image Pickup device            | 2/3-inch square pixel, 1.5 megapixel interline CCD |
| Color/Monochrome               | Monochrome   |
| Number of recording pixels · 2 | 1280x1024  |
| Quantization                   | 12 bits  |
| Sensitivity                    | Equivalent to ISO 800                              |

Table 4.1: Technical specifications of the DS-Qi1Mc camera that was used to acquire the images within the microscopy set-up.

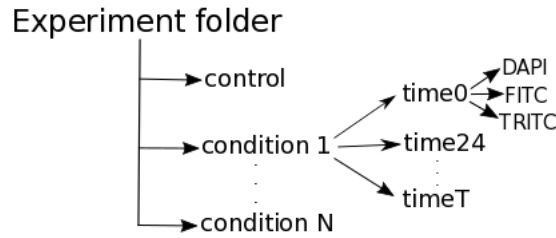


Figure 4.23: Folders structure required by the presented protocol for the correct analysis of the images. The parent folder represents the experiment and it contains one folder for each tested condition. Each one of these directories contains a number of folders equal to the time points tested for that condition. Each time point directory contains one folder for each fluorophore that contain the corresponding images.

#### 4.4.3 Fluorescence Quantification in Eukaryotic Cells

##### Inputs and outputs

As already described for the algorithm developed for the bacteria, also the one for eukaryotic cells requires the input images to be organized according to a precise scheme, that allows the program to automatically extract important information about the experiment and correctly associate each image to the corresponding sample and experimental condition. Specifically the images acquired within the same experiment must be divided according to experimental condition, time point and fluorophore, as exemplified in Figure 4.23.

After segmenting the images, the signal emitted by each cell is quantified, together with the cell density of each image and a number of output files are produced:

- *images.pdf*, that shows the segmented images in which all the channels have been combined.
- *graphs.pdf*, in which the changes in average cellular density and average fluorescent signals intensity, are shown over time, for all

the tested conditions. In this file are also plotted the fluorescence distributions for the tested fluorophores and their variation in time and with the experimental conditions.

- *ncells.txt*, in which are saved the numbers of cells segmented in each image divided by experimental condition and time point.
- *fluo\*.txt*, these files, one for each tested fluorophore, reports the recorded fluorescent intensities at single-cell level. Again values obtained for different conditions and time points are easily identifiable.

### Segmentation algorithm

As already described, the presented protocol takes advantage of the separate acquisition of the signal emitted by the different fluorophores to apply different elaborations to images with diverse characteristics. However it also makes use of the work-flow developed with the algorithm for bacterial cells both to compensate the distortions that affect a microscopy set-up and to correct the background of the acquired images. The former won't be discussed again as all the strategies previously described are integrated in this protocol without modifications. The latter, on the other hand, was adjusted to take into account the different size of the human cancer cells, with respect to the bacterial ones and different characteristics of the tested images. The background correction technique applied in both protocols is described in [110], it consists in subtracting each image to its gray-scale opening, obtained with a structuring element of the same size or bigger than the foreground objects. For images in which the signal is mostly in the cells' nuclei the kernel is square with a side of 100 pixels, while for proteins that localize in the cytoplasm or on the membrane the kernel's size is 150 pixels. Additionally, due to an higher autofluorescence of these images, the average intensity of the raw image is also subtracted.

The actual segmentation process is completely separate for images with different characteristics; the DAPI channel is segmented using a very similar procedure to the one developed for the bacterial images. The Zero Crossing edge detection method [91] is used to determine the nuclei edges and then the segmentation is completed by a hole filling procedure and some morphological operations. In this case to the detected edges is applied a binary closing, to prevent small gaps in the edges to interfere with the determination of the foreground objects, and only one binary erosion is executed on the resulting image. Furthermore the nuclei with an area smaller than 100 pixels are excluded from the

final image. Again, the identification of spurious edges is limited by filtering the images, prior to the segmentation, with a Gaussian kernel with standard deviation equal to 2.

Images in which the signal is mainly in the cytoplasm or on the membrane are segmented with the watershed algorithm, using as markers the nuclei segmented from the DAPI channel. This procedure is particularly effective in images where the foreground objects are not completely separate and can be exemplified as in Figure 4.17, where each marker is considered to be a water source, and edges are drawn when water from different water sources mix.

For proteins that are localized in the cell's nucleus the segmentation is not necessary, since the regions of interest have already been identified when the DAPI images were analysed. The signal emitted by each cell is simply quantified as the average intensity of the pixels in the regions segmented with the modified algorithm for bacterial cells.

### **Data analysis**

As previously described, fluorescent signals acquired with an optical microscope require two main adjustments: the data need to be normalized, to compensate for any confounding factor not considered during the calibration phase, and the number of cells considered for each condition must be equalized, to avoid introducing biases due to the different population's cardinalities.

In the analysis of the data presented in this section (Figures 4.19, 4.20) both these corrections were applied, using as a normalizer the average fluorescence intensity measured in the control condition at  $T=0$  h, and considering, for each condition, the largest possible population ( $\sim 70$  cells). When a larger number of cells was available the ones used for the analysis were randomly selected.

### **Experimental Protocol**

The protocol for the analysis of gene expression in eukaryotic cells was developed to elaborate images obtained with immunofluorescence assays. These experiments allow the detection of specific proteins through the use antibodies tagged with fluorescent molecules. The main steps of this protocol involve testing the experimental conditions of interest on cells seeded on coverslips and then fixing them with paraformaldehyde (3% in PBS), a chemical compound able to form covalent bonds between proteins in tissues, thus creating a snapshot of all the cellular processes at a given point in time. Successively the cells are treated with ethanol 100%, to induce the formation of pores in their

membranes and allow the entrance of the antibodies. These are then incubated with the cells for 30-45 mins, taking care to minimize the samples exposition to the light, due to the photo-sensitivity of the fluorophores. The same procedure is repeated for the DAPI dye and then the coverslips are mounted on a microscope slide and imaged.

In the specific experiment used to test the presented protocol two conditions were considered: in the treated one,  $\text{TGF}\beta$  was added to the culture media 24 hours after seeding, to give the cells enough time to adhere to the coverslips, while cells in the control condition were just monitored over time. The experiment lasted a total of two days and in the previous sections  $T=0$  h refers to the day after the seeding, when  $\text{TGF}\beta$  was administered to the cells.

## Chapter 5

# Cell invasiveness Quantification

### 5.1 Cell-Invasiv-O-meter

#### 5.1.1 Assay description and aim

Two of the hallmarks of cancer are an augmented invasiveness and an increase in the migratory capabilities of the cells. Thus assays that are able to quantify these characteristics are common tests to evaluate the efficacy of anti-cancer treatments or, in general, to compare different experimental conditions or cell lines.

One of the simplest frameworks to evaluate cell invasion is the so-called scratch wound healing assay [111]. It consists in creating a wound in a confluent cell culture, either by removing the cells in a specific area (using a pipette tip) or by preventing the culture from covering the entire plate using silicone inserts and then monitoring over time the cells' growth through bright field microscopy (Figure 5.1). This assay is very convenient, being easy to implement and not requiring specific equipment however, as highlighted in [112], it lacks standardization and thus its results are scarcely reproducible. There are several reasons behind this concern, but one of them is the rather simplistic way of quantifying the result of these experiments. The two main strategies involve either measuring the time required to completely close the wound or evaluating the gap's width at specific time points. The second method is generally preferred since it allows to directly compare different experimental conditions and associate the invasiveness of a cell population to other variables measured at the same time points (e.g. the level of expression of specific genes). The major problem with this technique is that the gap width varies considerably along the wound and thus taking a single measurement in a randomly chosen point is

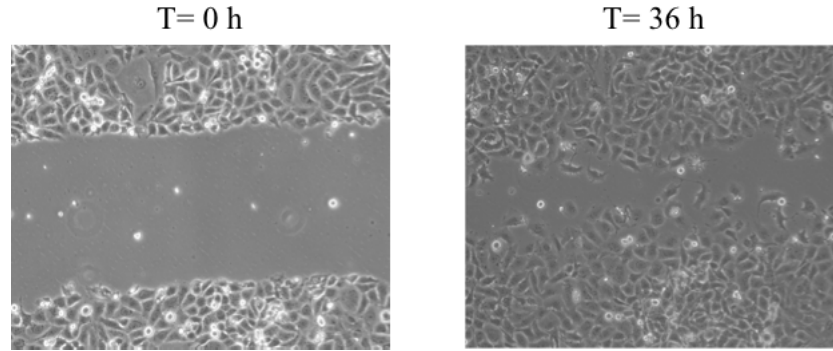


Figure 5.1: Exemplification of a scratch wound healing assay. A wound is created in a confluent cell culture and then the rate of closing of the gap is monitored with an optical microscope.

very unlikely to be representative.

A more appropriate strategy involves quantifying the gap area, but this can be difficult to achieve with standard image processing softwares, like ImageJ [113] or Photoshop, in which the area selection tool is a regular polygon that is ill suited to capture the irregularities of the wound.

To address these limitations in 2009 was presented TScratch [114], a semi-automated tool that uses an edge detection algorithm based on the discrete curvelet transform to segment the wound and quantify its area. This software is freely available and has a simple graphical user interface (GUI) that allows the user to revise the processed images and to change the algorithm's parameters to ensure the correct identification of the cell free area. Despite being user friendly and well documented TScratch hasn't become the standard for the analysis of scratch wound healing assays, probably in part due to the complexity of the curvelet transform and the unintuitive connection between the value of the parameters and their effect on the images.

In the following, Cell-Invasiv-O-Meter will be presented, it is a tool developed for the analysis of scratch wound healing assays, that segments the wound, evaluating the local entropy of the images, then quantifies the cell free area and produces a bar graph representing the results normalized to the values of the provided control. The fully automatic process make this software ideal for analysing a large number of images, furthermore the immediate connection between the tunable parameters and their effect on the result improves its usability.

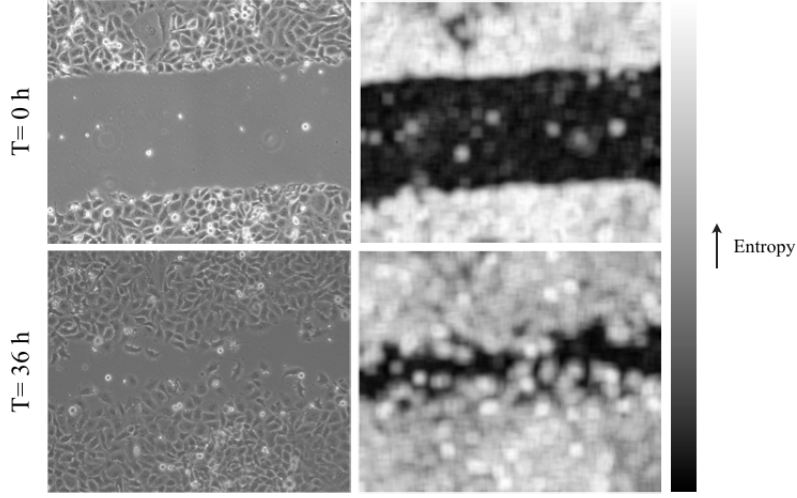


Figure 5.2: Local entropy images associated to the images of Figure 5.1

### 5.1.2 Algorithm development

Cell-invasiv-o-meter is a function implemented in Matlab R2012a [81] that identifies the cell free regions of the image computing the local entropy (Equation 5.1), where  $M$  is the number of gray levels of the image (256 for 8 bits images) and  $p_k$  is the probability of gray level  $k$ .

$$H = - \sum_{k=0}^{M-1} p_k \log_2(p_k) \quad (5.1)$$

Entropy was introduced in the field of Information Theory by Shannon in 1948 [115]. It quantifies the information encoded in a message as related to the probability of receiving it. Thus messages or events that are very likely to be received/occur, like getting head flipping Two-Face's coin, will carry little information and thus will have low entropy (zero in the proposed example). On the other hand unexpected events or messages will be associated to a high entropy due to their low probability.

Computing the local entropy of an image corresponds to creating another image, of the same size, in which the value of each pixel is the entropy computed over the neighbourhood of the pixel itself. Thus highly homogeneous regions, like the cell free areas, will have low entropy, while the parts of the image containing the cells will be associated to a much higher entropy. This is exemplified in Figure 5.2, where dark shades are associated to low entropy, while lighter colors identify regions with high values of this parameter. This transformation clearly isolates the cell free area, even excluding the small cell isles left behind

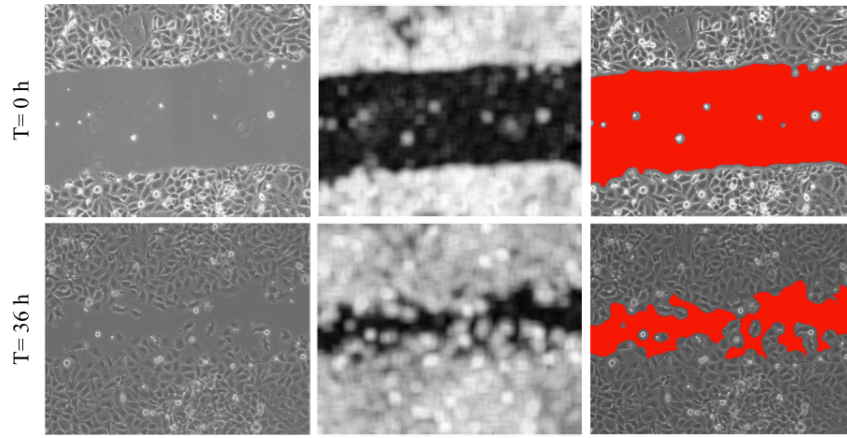


Figure 5.3: Segmentation's results showing in red the area recognized as the wound by Cell-Invasiv-O-Meter.

when creating the wound.

The next step of the analysis involves segmenting the local entropy image using a global threshold identified with the Otsu's method [116]. In Figure 5.3 the result of the segmentation is presented, the wound (identified in red) is correctly captured even at  $T=36$  h where the cell free area is highly irregular and the channel is effectively divided in two regions by a small group of cells that have completely bridged the gap. The cell free area is quantified in Cell-Invasiv-O-Meter as the number of pixels belonging to the wound (the red region in Figure 5.3). Even though it would be possible to convert it to  $\mu M^2$ , knowing the magnification of the objective used to acquire the pictures, it didn't seem necessary since the results of scratch wound healing assay are generally presented as relative to an appropriate control, thus making them independent on the unit of measure. Furthermore, since the conversion would be between number of pixels and  $\mu M$ , it would be necessary to assign a regular shape to the wound and compute the equivalent area, to obtain a surface measure.

The results of the elaboration are saved by Cell-Invasiv-O-Meter in a .mat file that can be read by another function, named compare-Cell-Invasive-O-Measures that computes the area decrease for each condition and time point with the respect to the initial wound area, averages the replicates and relates it to the values obtained for the provided control. At the end of the analysis this function produces a histogram that summarizes the experiment's results (Figure 5.4).



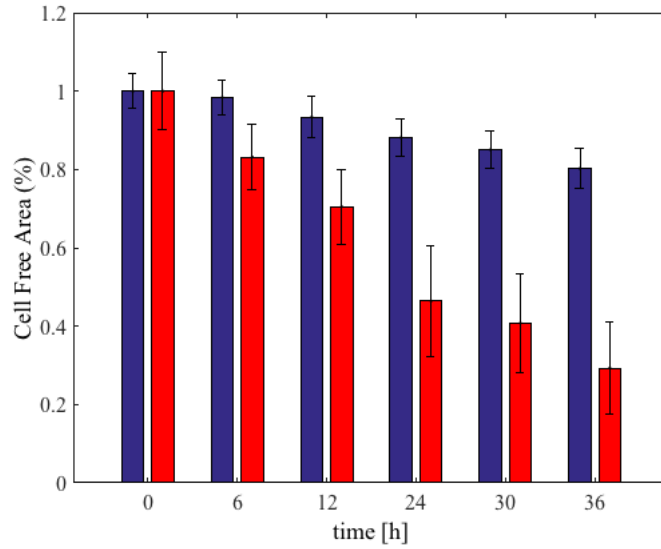


Figure 5.4: Results of the scratch wound healing assay analysed with Cell-Invasiv-O-Meter. The cell culture supplemented with TGF $\beta$ , represented in red, is decisively more invasive than the control (reported in blue), covering more than 60% of the wound in 36 h.

### 5.1.3 Results

Cell-Invasiv-O-Meter was used to analyse a simple scratch wound healing assay in which a cell population whose media was supplemented with TGF $\beta$  was compared to an untreated control over a time period of 36 hours. The presence of TGF $\beta$  is expected to increase cellular invasion, being one the most effective inducers of Epithelial to Mesenchymal transition (EMT) *in-vitro*. In Figure 5.4 are reported the results of this experiment. The control population, represented in blue, is significantly less invasive, with a decrease in cell free area at 36 h that is below 20% with respect to the initial wound area. Adding TGF $\beta$  to the culture media significantly enhances the invasive capabilities of the cells that, in the same time, are able to cover more than 60% of the original cell free region.

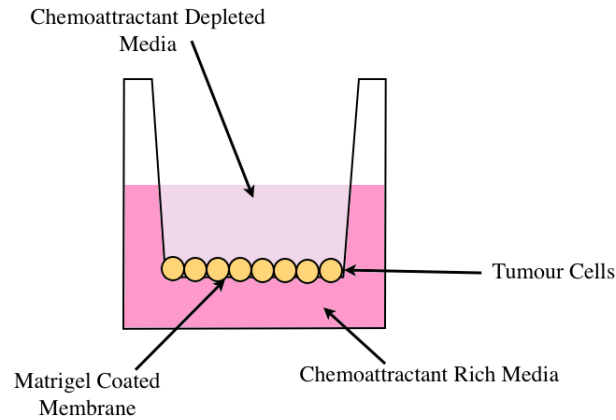


Figure 5.5: Schematic representation of a transwell assay. Cells are seeded on a Matrigel coated membrane in the transwell insert. A chemoattractant induces the migration of the cells toward the bottom of the membrane. The number of cells that will be able to degrade the Matrigel and cross the membrane is proportional to the invasiveness of the population.

## 5.2 I-AbACUS (Invasion-Assay Assisted cell CoUnting Software)

### 5.2.1 Assay description and aim

While being simple and inexpensive the scratch-wound healing assay is a very simplified representation of the conditions in which the cancer cells migrate and invade *in-vivo*. Cells are seeded on Petri dishes that, being made of polystyrene, create an environment that is considerably different from a biological tissue. Furthermore scratch wound healing assays model migration and invasion as 2D processes. Beside being far from reality this simplification introduces an important confounding factor, that is the cells' growth rate. Indeed the invasive capabilities of cells that have a short duplication time will be overestimated, as it is not possible to differentiate between the decrease in cell-free area due to migration of the cells or due to the population's growth.

For these reasons migration and invasion are generally quantified *in-vitro* using transwell assays (Figure 5.5). In these experiments a defined number of cells is seeded on the upper layer of a cell permeable membrane that, when studying invasion, is coated with Matrigel, a polymer that simulates the extracellular matrix. A chemoattractant, generally a higher nutrients concentration, attracts the cells toward the bottom of the membrane and induces their migration.

The invasiveness of a population is quantified as the number of cells that are able to cross the membrane within a specified time frame.

This result is generally obtained counting manually the cells, either from images acquired with an optical microscope, using image processing software with object counting functionalities [117, 118], or directly from the instrument, with the help of a mechanical cell counter [119, 120, 121, 122]. This introduces a strong dependence on the user's ability to correctly identify the cells, task that might be quite complex when each field contains a significant amount of cells and/or they can assume a range of phenotypes. Furthermore the experimental protocol can vary significantly in the number of fields considered for each membrane, in the magnification of the objective used to image the cells and in the initial cell density. All these factors limit the accuracy and reproducibility of the results.

To address these limitations I-AbACUS was developed; it is a custom made software that guides the analysis of images acquired during transwell migration/invasion assays. By applying the same procedure to every image, it improves the results' reproducibility and by segmenting and automatically classifying the cells it reduces the dependence on the operator and its experience with the assay.

### 5.2.2 Description of I-AbACUS

I-AbACUS was developed to integrate all the steps of the analysis of transwell assays, through its simple and intuitive graphical user interface (GUI). However its main features are the segmentation of the images and the identification of the cells through the classification of the foreground regions. This second step can be executed either using a filter empirically determined (EF) or with a trained Support Vector Machine (SVM). They both consider three characteristics for each foreground region: (i) its area, that allows the exclusion of debris and small irregularities, (ii) the interquartile range of the saturation of its pixels, that is a measure of dispersion of the pixel's values and can be used to prevent out of focus cells to be recognized and (iii) its circularity, computed as in Equation 5.2 where A and P represent, respectively the area and the perimeter of the object.

$$C = \frac{4\pi A}{P^2} \quad (5.2)$$

This last parameter allows to exclude from the analysis the membrane's pores by comparing the circularity of the segmented region to that of a circle.

When the EF is used, a foreground region is classified as a cell if at least two of the previously mentioned characteristics have values within predefined ranges. This method is versatile and generally applicable to

cell lines with different phenotypes, but the classification of a specific type of cells can be improved through the use of a SVM, a supervised learning method that, provided an adequate training set, identifies a classifier able to distinguish between cells and non-cells using the three morphological characteristics previously described. The training set, required to train the SVM, can be produced analysing images of the specific kind of cells with the EF, as the program will automatically save a text file containing the morphological characteristics of the segmented objects and their classification.

In I-AbACUS cells are segmented applying the marker-controlled watershed transform to the saturation channel of the images coded in the HSV colorspace. This algorithm [123] is particularly effective when the foreground regions are contiguous and consists in placing metaphorical water sources on the local minima of the gradient of the image and drawing edges when water from two adjacent basins meets (Figure 4.17). To avoid the segmentation of spurious minima the watershed procedure is preceded by the marking of each foreground and background region. In I-AbACUS the foreground markers are obtained through an opening by reconstruction and a closing by reconstruction, followed by a procedure that removes the smallest markers. Figure 5.6 exemplifies this process using a black and white image for simplicity. The erosion of the original image (Figure 5.6 **d.**) identifies all those foreground regions that match the employed structuring element and its subsequent reconstruction, using the original image (Figure 5.6 **a.**) as a mask (Figure 5.6 **e.**) leads to an image in which only the letters t and f are present. This is due to the fact that only those letters contain the pattern defined in the structuring element, a disk of radius 2 pixels. The reconstructed image is then dilated, using the same structuring element and complemented (Figure 5.6 **b.**); this image is then reconstructed, using as a mask the complemented version of Figure 5.6 **e.** obtaining the complement of the final result (Figure 5.6 **c.**). This procedure removes small imperfections in the image while maintaining the original shape of the cells, furthermore it creates a flat maxima within each region allowing the identification of the foreground markers through the computation of the regional maxima.

The background markers are identified applying a coarse segmentation to the reconstructed image and computing the skeleton by influence zone of the background. This procedure ensures a good separation between foreground and background markers through the identification of the ridge lines of the watershed transform computed on the distance transform of the segmented image. Figure 5.7 shows the result of this step applied to the black and white image used in Figure 5.6. The ridge

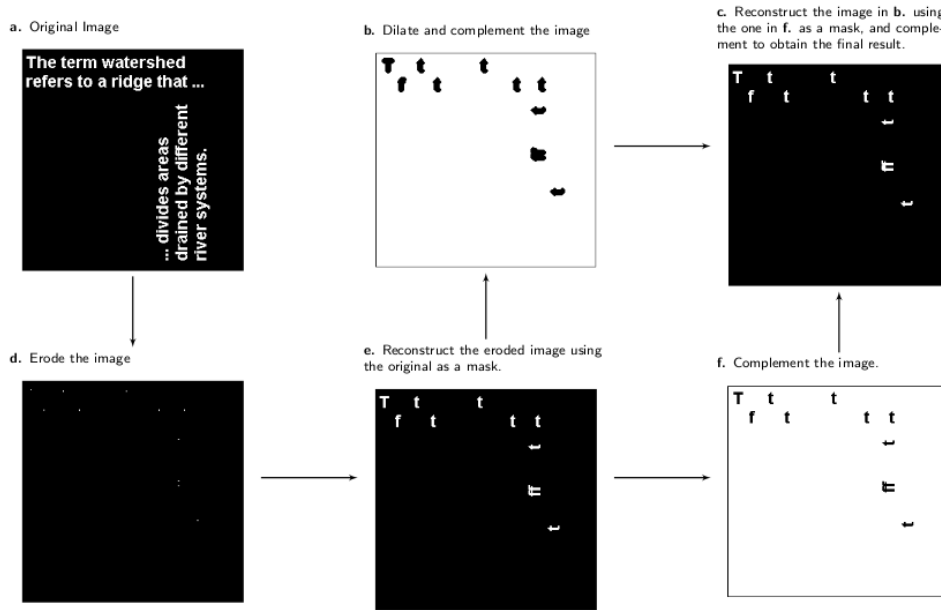


Figure 5.6: Schematic representation of the procedure implemented in I-AbACUS to determine the foreground markers. **a.** The original image, in this example for simplicity a black and white one is used. **d.** Result of the morphological erosion of the original image using as a kernel a disk of radius 2 pixels. **e.** Reconstruction of the eroded image using the original image as a mask. In this example only *ts* and *fs* contain the pattern chosen as structuring element and thus are the only ones that remain after the erosion. **b.** Dilation and complementation of the reconstructed image. **f.** Result of the complementation of the reconstructed image. **c.** Final image on which the foreground markers are computed, as the regional maxima. It is obtained complementing the reconstruction of the image in **b.** using the image in **f.** as mask.

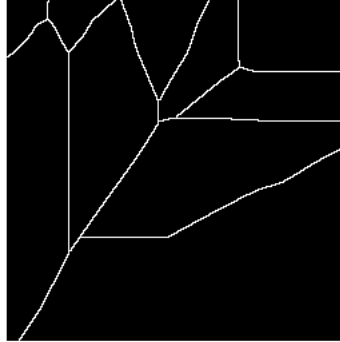


Figure 5.7: Result of the Skeleton by influence zone that leads to the definition of the background markers. The ridge lines identify 10 different regions, one for each foreground marker.

lines divide the image in 10 sectors, one for each foreground marker, thus ensuring their spatial separation.

Prior to image segmentation, the background of the images is uniformed, by masking the current image with its coarse segmentation obtained applying the Otsu's method [116].

After the automatic analysis of each image, the user can modify the proposed cell count by selecting sequentially each area not correctly segmented/classified and either selecting one of the four possible alternatives proposed by I-AbACUS or directly providing the number of cells in the specified region. Each alternative is produced using a different dimension for the kernel used during the opening and closing by reconstruction, leading to the recognition of either smaller or larger cells. To minimize the elaboration time and eliminate any delay in this phase of the analysis, the five alternative segmentations are implemented as a parallel process.

The cell count is automatically updated by the program and at the end of the analysis the results can be visualized, in tabular form and as in histogram in which the average and standard deviation of the cell count for each replicate and condition is reported. Finally they can be saved as an excel file for further elaboration.

### 5.2.3 Use of the learning algorithm

SVMs are linear classifiers that identify the hyperplane that classifies the provided data with the largest margin (Figure 5.8). Formally it consists in the constrained minimization of Equation 5.3 where  $\beta^T x + \beta_0$

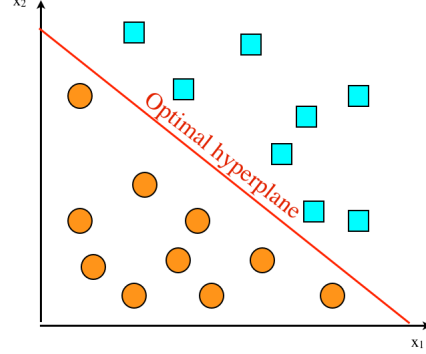


Figure 5.8: Exemplification of the classification strategy used by SVMs. It consists in identifying the equation of the hyperplane that maximizes the distance between the samples belonging to different classes. In this example only two characteristics are considered, thus the hyperplane is a line, but this strategy is applicable to arbitrary large parameter spaces. The only condition is that the two classes must be linearly separable, if this condition is not satisfied the kernel method must be applied.

is the hyperplane,  $\beta$  a weight vector,  $\beta_0$  is the bias.

$$\min_{\beta, \beta_0} L(\beta) = \frac{1}{2} \|\beta\|^2 \text{ subject to } y_i(\beta^T x_i + \beta_0) \geq 1, \forall i \quad (5.3)$$

The imposed constraints require the hyperplane to correctly classify each training example ( $x_i$ ), since  $y_i$  represent the label associated to each sample.

To make this approach generally applicable, even to non-linearly separable classes, I-AbACUS implements the kernel method. It consists in using a polynomial kernel ( $\phi$ ) to map the original input space in a feature space of higher dimensionality in which the data are linearly separable and then computing the maximum margin hyperplane (Figure 5.9).

Being supervised learning algorithm, SVMs require the user to provide a set of examples to identify the hyperplane. In I-AbACUS it can be built by using the EF to analyse images of the specific kind of cells that the user wants the trained SVM to recognize. Indeed the program automatically saves a .mat file, in a folder named Results in the same path as the code, containing the morphological characteristic (area, circularity, saturation's IQR) of each segmented region and their

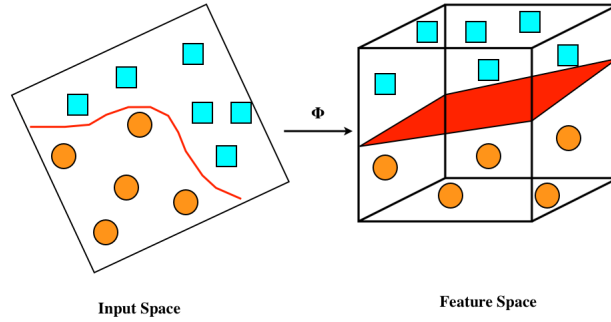


Figure 5.9: Graphical representation of the kernel method used to make SVMs applicable to non-linearly separable datasets. A polynomial kernel ( $\phi$ ) is used to transform the input space in the feature space. The latter has an higher dimensionality, than the input space and allows for the linear separation of the two classes.

classification as cells (1) or not-cells (0).

These files can be used as input in the learning algorithm tab of I-AbACUS, that is responsible for running the function that minimizes Equation 5.3. Once the model has been identified it is saved in a .mat file, that can then be used, while analysing an invasion assay with I-AbACUS, to improve the classification of a specific type of cells.

#### 5.2.4 Results

I-AbACUS was validated through a series of experiments that compared its results to the ones obtained with the traditional technique (manual count of the cells using ImageJ). Beside demonstrating the equivalence of the two methods over a wide dynamic range, some of the most important limitations of the standard analysis technique of invasion assays, were evaluated to determine if the use of I-AbACUS could address them. In the following the results of this analysis are presented.

##### Equivalence between I-AbACUS and the traditional analysis

To demonstrate that I-AbACUS's results are equivalent to those obtained manually counting the cells with the aid of ImageJ [113], the



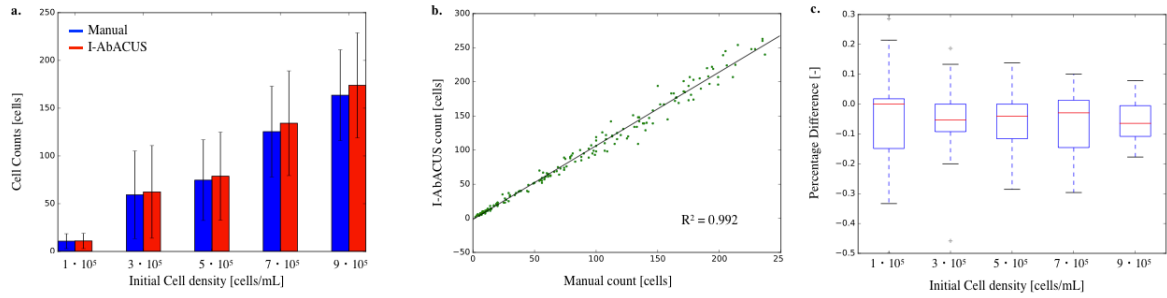


Figure 5.10: Comparison between the cell counts measured with I-AbACUS and the ones obtained manually. **a.** Average cell count for every initial cell density. The results of the manual count are reported in blue, while red identifies the ones obtained with I-AbACUS. **b.** Correlation plot between the counts obtained with both instruments on single images ( $R^2 = 0.992$ ). **c.** Percentage difference between the manual and I-AbACUS's cell counts as a function of the initial cell density.

same images (180) were analysed with both methods by the same expert operator. For this comparison a cell line named OVCAR4 was used, it is an high grade ovarian serous adenocarcinoma that shows high morphological variability and is widely used in the study of ovarian cancer. To obtain images with different cell densities, five distinct initial concentrations of OVCAR4 cells were seeded in the Matrigel coated transwells and invasion was determined after 48 h, both in terms of cell counts and counting times.

The cell counts show a very good agreement both considering the average cell count for each condition (Figure 5.10 **a.**) and single images (Figure 5.10 **b.**), where the Pearson's correlation coefficient between the manual and the I-AbACUS counts is 0.992. The errorbars in Figure 5.10 **a.**, that represent the standard deviation of the cell counts for the corresponding condition have comparable amplitudes for the two methods. The percentage difference, computed considering the manual count as the theoretical value, doesn't considerably change between the tested conditions, thus demonstrating the equivalence of the two methods on a wide range of cellular densities (Figure 5.10 **c.**).

In Figure 5.11 **a.** the average counting time is reported as a function of the initial cell density. The analysis with I-AbACUS (reported in red) takes consistently longer than the manual one, partly due to the fact that the time required by the program to elaborate the images is included in the measure, thus introducing a non-negligible offset. This is also represented in Figure 5.11 **b.** where the dependence of the counting time on the number of cells in the image is considered.

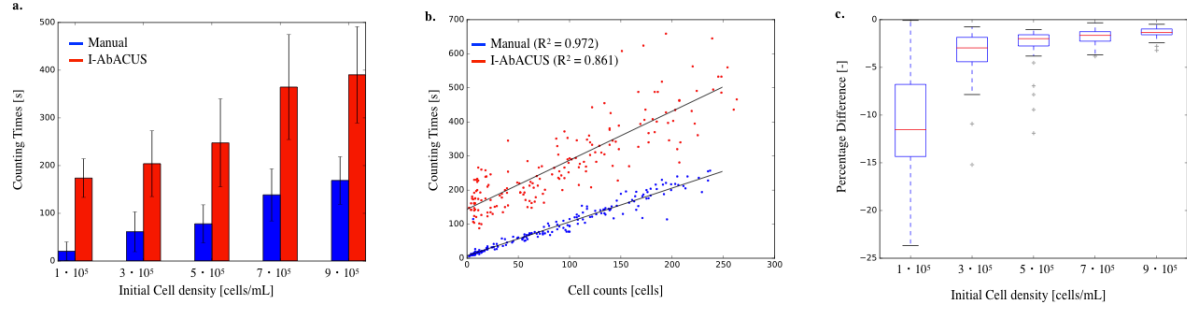


Figure 5.11: Comparison between the counting times recorded with a stopwatch while collecting the data reported in Figure 5.10. **a.** Average counting time as a function of the initial cell density. The blue bar represents the manual counting time while the red one the I-AbACUS result. **b.** Relation between counting time and the corresponding cell count on single images. The red dots, representing the I-AbACUS results, show a lower dependence on the image density ( $R^2=0.861$ ), while the number of cells counted manually for each time unit is approximately constant ( $R^2=0.972$ ). **c.** Percentage difference between the counting times, as a function of the initial cell density.

The traditional analysis shows a very good correlation between these two variables ( $R^2=0.972$ , blue dots), showing how the number of cells counted per time unit was almost independent on the total number of cells in the image. On the other hand the I-AbACUS counting times showed a lower dependence on the image density ( $R^2=0.861$ , red markers) and an offset of about 120 s. The exclusion of the elaboration time, however, would have reduced the accuracy and the correctness of the comparison, since the image's elaboration is an integral part of the I-AbACUS workflow. Figure 5.11 **c.** reports the percentage difference between the counting times recorded with the two methods. It markedly decreases with the initial cell density (almost 10 fold comparing  $1 \times 10^5$  and  $9 \times 10^5$  cells/mL), showing how the inconvenience of a non-negligible elaboration time can be compensated, exploiting the lower dependence of the I-AbACUS counting times on the cell density. Since the percentage difference on the cell count does not depend on the initial cell density (Figure 5.10 **c.**), considering images with an higher number of cells would increase the efficiency of the analysis, reducing the incidence of the elaboration phase on the total counting time, while increasing the robustness of the results, due to the increase in population size.

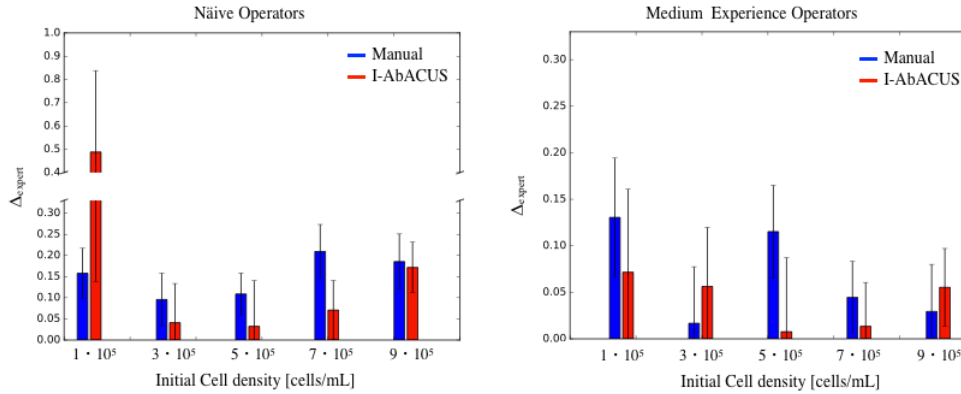


Figure 5.12: Evaluation of the inter-operator variability. **a.** Results of the Naïve operators reported as the difference between their cell counts and the ones obtained by an expert operator. All the data are normalized with respect to the results of the I-AbACUS expert. **b.** Difference between the results of the Medium experience operators and that of the expert as a function of the initial cell density. Again all the cell counts are normalized with respect to the corresponding cell counts of the I-AbACUS expert.

### Inter-Operator Variability

One of the most critical aspects of the traditional analysis of transwell assays is the high inter-operator variability, that is the results are significantly dependent on the operator's ability to correctly identify the cells.

To test if the use of I-AbACUS could address this limitation five additional operators were asked to count two sets of 22 and 23 images respectively, extracted from the one used for the previous comparisons and covering all the tested initial cell densities. One operator for every level of experience was assigned to the first dataset while the other naïve and medium experienced operators counted the second group of images. They were asked to count the cells in the provided images both with I-AbACUS using the EF and with the traditional approach. Furthermore to eliminate any possible confounding factor all the users were given the same set of instructions on how to count cells, both manually and with I-AbACUS. Figure 5.12 shows the results of this test as difference between the average cell counts obtained by the naïve and medium operators and that of the expert (Figure 5.12 **a.** and **b.** respectively). All the data were normalized with respect to the values obtained by the I-AbACUS expert operator, that produced the results of the previous comparisons.

The use of I-AbACUS reduces, in most cases, the difference between

the counts obtained by the expert operator and those of less skilled users, showing how being provided with a suggested result reduces the inter-operator variability and the dependence of the results on the user's experience. This is particularly evident when looking at an initial cell density of  $5 \times 10^5$  and  $7 \times 10^5$  cells/mL, that can be considered to be the target cellular density for transwell invasion assays with OVCAR4 cells. The error of the experienced operator will be minimum, since this is the type of images on which she has more training, while the other operators will find them challenging due to the increased number of cells.

### Intra-operator Variability

Intra-operator variability refers to the difference in the number of cell counted by the same operator, when repeating the analysis of an image. Low intra-operator variability is a desirable characteristics as it is associated to the assay's repeatability and the robustness of the results. To evaluate if the use of I-AbACUS improve this aspect, one operator for each level of experience counted again a subset of images (9) from the ones used in the inter-operator variability test. In this case only three initial cell densities were considered ( $1 \times 10^5$ ,  $5 \times 10^5$ ,  $9 \times 10^5$  cells/mL) and the operators were asked to execute the analysis as in the previous test.

Figure 5.13 reports the results of this test as a scatter plot in which the percentage difference between the two counts obtained for the same images are reported. The use of I-AbACUS significantly reduces the intra-operator variability, measured as the median percentage difference between the two counts (Table 5.1). In this test the Naïve opera-

|                 | Manual Count | I-AbACUS Count | Marker          |
|-----------------|--------------|----------------|-----------------|
| Naïve Operator  | 11.1%        | 0%             | cyan circle     |
| Medium Operator | -49.3%       | 25%            | magenta square  |
| Expert Operator | -2.9%        | 1.3%           | yellow triangle |

Table 5.1: Median intra-operator variability measured for each user and counting method.

tor did not modify the cell count proposed by I-AbACUS thus leading to exactly the same result for every image. Furthermore the intra-operator variability seems to have a lower dependance on the level of experience of the operator, since the results of the medium experience operator are less robust than the ones of the Naïve user.

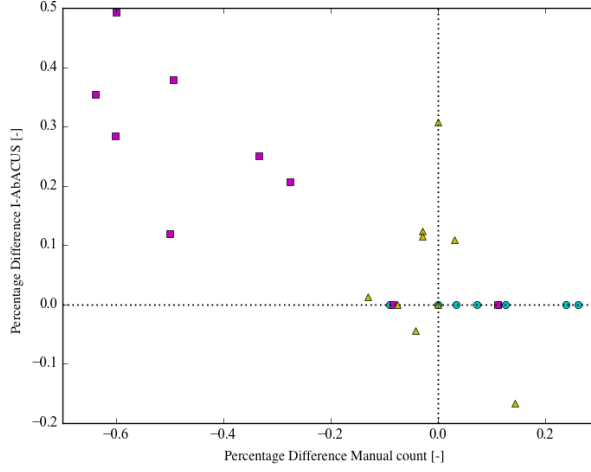


Figure 5.13: Analysis of the intra-operator variability. The percentage differences between two consecutive counts of the same images with both methods are reported. Each symbol represents a different user, the cyan circles correspond to the Naïve user, the magenta squares to the Medium experience user and the yellow triangles to the Expert.

### Versatility

To test the ability of I-AbACUS to analyse images containing cells other than OVCAR4, a dataset of 16 images was acquired using A2780 cells, another ovarian cancer cell line that shows a less diverse phenotype and cells that are smaller and rounder, when compared to OVCAR4s (Figure 5.14). For this test only the optimal initial cell density of  $5 \times 10^5$  cells/mL was used and the same experienced user counted the all images both with the traditional approach (in blue in Figure 5.15 **a.**, and 5.15 **c.**) and with I-AbACUS, using the EF as classification method.

Figure 5.15 reports the results of this analysis that closely mirror the ones in Figures 5.10 and 5.11, the average cell counts are equivalent (Figure 5.15 **a.** percentage error 20%) and the correlation between the results obtained on the single images is very high,  $R^2=0.975$  (Figure 5.15 **a.**).

The counting time is still significantly lower when the images are analysed manually but again its dependance on the image density is lower with I-AbACUS (Figure 5.15 **c.**,  $R^2_{I-AbACUS}=0.862$  and  $R^2_{Manual}=0.980$ ) hinting to the possibility of incrementing the initial cellular density to increase the efficiency of the test and the robustness of the results, as shown for the OVCAR4 data.

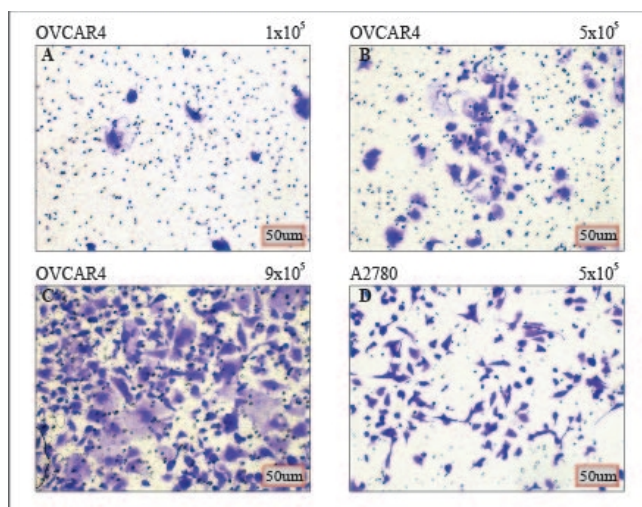


Figure 5.14: Comparison between some of the different cell densities and phenotypes used for the validation of I-AbACUS. Panels **A**, **B** and **C**. represent OVCAR4 cells at three different initial cell densities ( $1 \times 10^5$ ,  $5 \times 10^5$  and  $9 \times 10^5$  cells/mL), while in panel **D**. a representative image of A2780 cells (initial cell density  $5 \times 10^5$  cells/mL) is reported. Figure kindly realized by Claire Henry PhD.

Being able to elaborate, using the same parameters, images featuring cells with different morphologies, greatly expands the I-AbACUS applicability, not only to different cell lines but also to the analysis of invasion assay that use treatments that modify the morphology of the cells.

### Comparison between EF and SVM

To compare the two cell classification methods implemented in I-AbACUS the same dataset used to demonstrate the equivalence with the manual count was analysed again by the same operator, using a SVM trained on a completely independent set of images. The training set was composed of 160 images, 48 of which were acquired from membranes on which A2780 cells were seeded, while the others were obtained from experiments involving OVCAR4. All the images were counted both manually and with I-AbACUS obtaining a normally distributed percentage difference, that was less than 35% in 90% of the images (Figure 5.16). The relation between the two cell counts is shown in Figure 5.17, it is linear and the correlation coefficient is 0.980.

Figure 5.18 reports the results of the comparison between EF and SVM as the percentage difference between the results obtained with

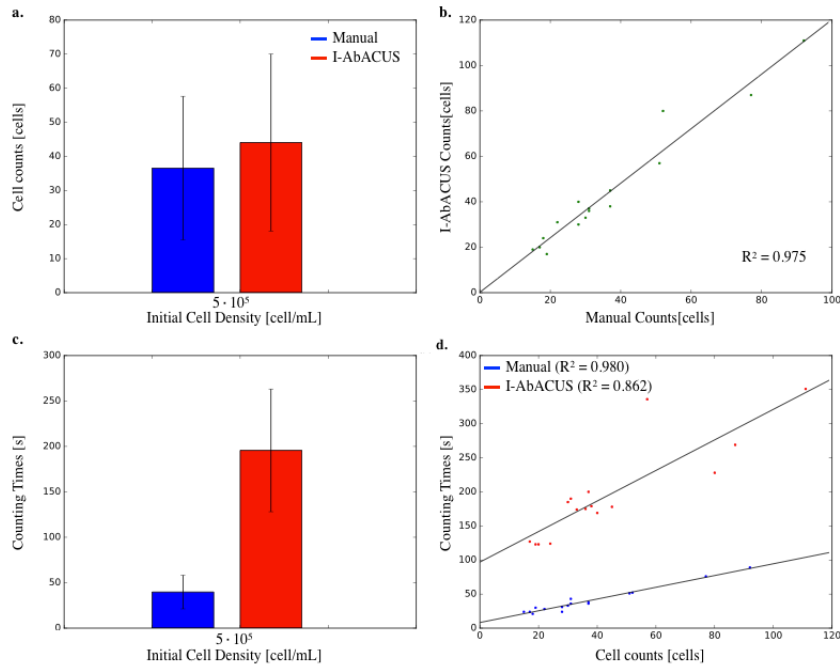


Figure 5.15: Analysis of the experiment with the A2780 cells, aimed to evaluate I-AbACUS's ability to analyse cells with different morphologies. **a.** Comparison between the average cell counts (and the corresponding standard deviations). In blue is reported the value obtained with the manual count while red identifies the I-AbACUS result. **b.** Correlation plot comparing the results obtained on single images ( $R^2 = 0.975$ ). **c.** Comparison of the average counting time measured when obtaining the results shown in **a.**. **d.** Relation between counting time and corresponding cell count ( $R_{Manual}^2 = 0.980$ ,  $R_{I-AbACUS}^2 = 0.862$ ).

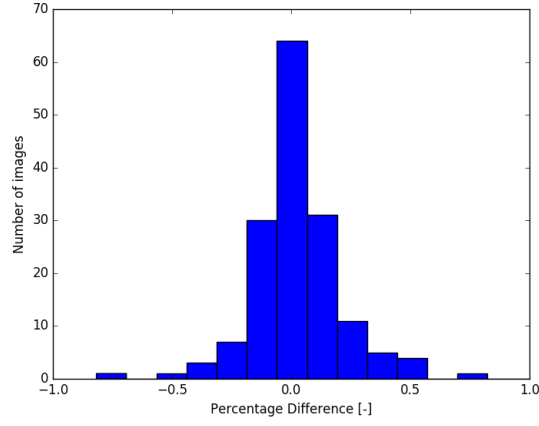


Figure 5.16: Histogram of the percentage difference between the I-AbACUS cell count and the manual one computed on the training set.

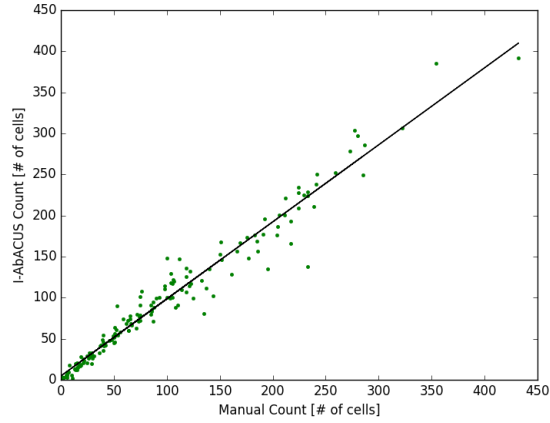


Figure 5.17: Correlation plot between the cell counts obtained with I-AbACUS and the traditional technique on the image of the training set ( $R^2=0.980$ ).



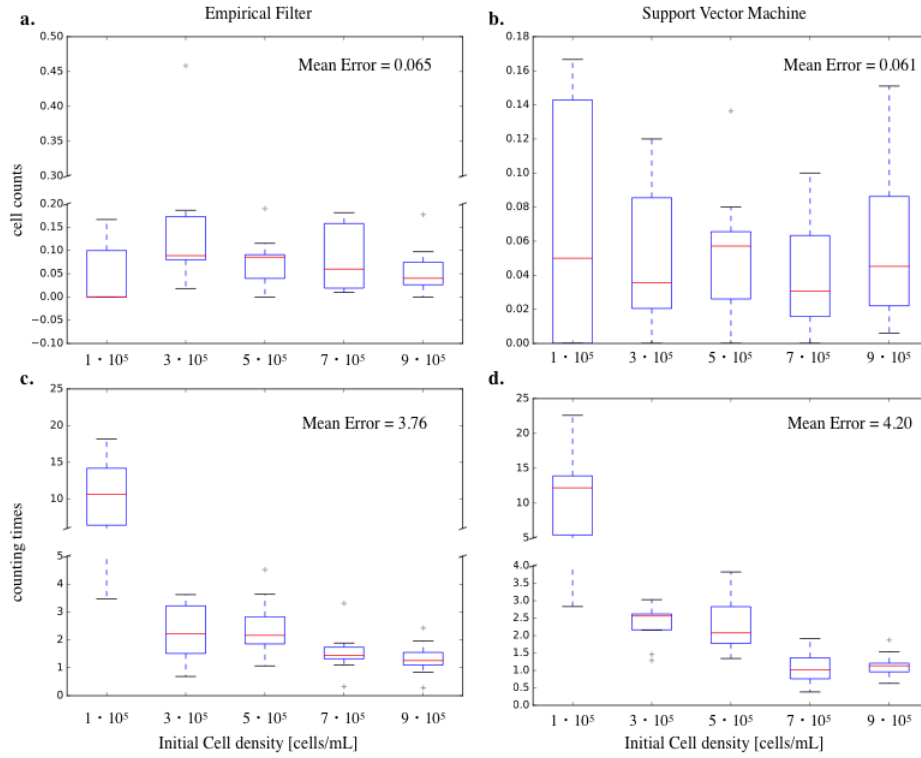


Figure 5.18: Comparison between the two cell selection methods implemented in I-AbACUS. **a.** Percentage difference between the cell counts obtained with the EF and the manual count as a function of the initial cell density (average difference 0.065). **b.** Relation between the percentage difference of the cell counts obtained using the SVM and the manual count as a function of the initial cell density (mean error= 0.061). **c.** Percentage difference between the counting times obtained when obtaining the data in **a.** **d.** Same as in **c.** but using the counting times measured when using the SVM.

I-AbACUS and the manual count. The use of the SVM doesn't significantly affect the counting time (Figure 5.18 **c.,d.**) or the average cell count (Figure 5.18 **a.,b.**), however it increases the reliability of the count reducing the number of outliers and their distance from the population.

### 5.3 Discussion

The two softwares presented in this chapter were developed to analyse the results of the two most commonly used techniques for the *in-vitro* evaluation of cell migration and invasiveness.

The first one considered the scratch wound healing assay, a 2D test has the advantages of being simple to implement and not requiring specific instrumentation. In this case local entropy and a simple thresholding procedure were used to automatically separate the regions of the image containing the cells from the empty one. The quantification of the latter provided an accurate measure of the rate at which the considered population was filling the gap artificially created in a confluent culture. Preliminary tests, executed using this tool, showed its ability to distinguish between alternative experimental conditions and improve the reliability and accuracy of the results (manuscript in preparation).

Despite the accuracy of this test, the bidimensional environment represents a significant simplification of cell migration, since it is not able to compensate for cell doubling and does not include the interaction between the cells and their environment. Transwell assays are able to address these limitations, studying cell migration in a 3D environment that includes a polymeric layer designed to simulate extracellular matrix. The images acquired during this *in-vitro* assay can be analysed using the second tool presented in this chapter. I-AbACUS was extensively tested and its equivalence with the standard analysis technique for this experiment was demonstrated. Furthermore the software-aided analysis was shown to be less dependent on the operator's experience and more repeatable, leading to an improvement of the quality of the result.

The evaluation of cell migration and invasiveness fundamental for the characterization of cancer cell lines and determining the effectiveness of potential treatments. Despite their importance their evaluation is often qualitative or semi-quantitative, leading to results that lack accuracy and might be affected by significant biases. The two computational tools here described contribute to addressing these limitations, introducing standard protocols for the analysis of the images acquired

during the considered assays and suggesting guidelines for improving the experiment's repeatability.

## 5.4 Materials and Methods

### 5.4.1 Cell-Invasiv-O-meter

#### Scratch Wound Healing assay

The scratch wound healing assay is an *in-vitro* experiment designed to evaluate the invasiveness of a cell culture in a simplified 2D setting. It consists in creating a scratch in a monolayer of cells and quantifying the closing rate of the wound.

The results presented in the previous section were obtained by Alice Pasini PhD. from an experiment in which an untreated control was compared to the treatment with TGF- $\beta$ , using the A549 a lung cancer cell line. This treatment is expected to increase the invasiveness of the cells by inducing the EMT.

A pipette tip was used to create the wound, and two thin lines were drawn, perpendicularly to the wound with a permanent marker (Figure 5.19). These marks on the bottom of the plate allows to univocally identify four regions of the wound, shown as shaded squares in Figure 5.19 that were monitored every 6 hours for a period of 36 hours. The plates were incubated at  $37^{\circ}C$  and 5%  $CO_2$  and, at each considered time point, images of the four regions of the wound adjacent to the marks were acquired with an optical microscope. The 48 images obtained during this experiment were analysed with Cell-Invasiv-o-meter and the results are presented in Figure 5.4.

#### Segmentation of the Wound and Cell Free Area Quantification

Cell-Invasiv-O-Meter is a freeware function downloadable at [124], that was developed in Matlab R2012a [81]. It requires the Image processing toolbox and the function Imoverlay, developed by Matt Smith and freely available at [125].

To identify the cell free area this algorithm exploits the local entropy (Equation 5.1), a measure of the information stored within the neighbourhood of each pixel. The wound, being approximately uniform in color, will have a significantly lower entropy, than the rest of the image (Figure 5.2), allowing its identification with a simple global thresholding procedure (Otsu's algorithm [116]).

This algorithm divides the pixels of the image in two classes minimizing the intraclass variance or, equivalently, maximizing the inter-class

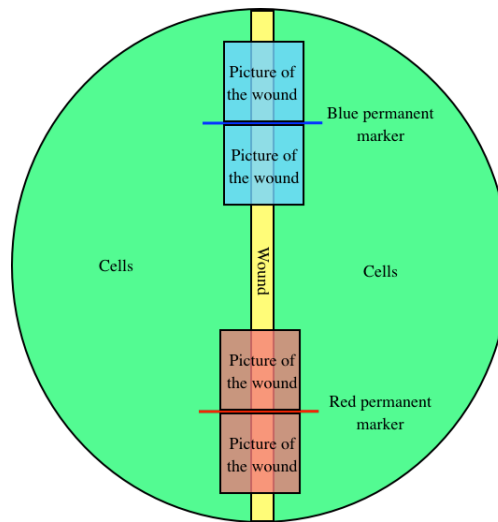


Figure 5.19: Representation of the scratch wound healing assay that was used to compare the effect of the treatment with  $\text{TGF}\beta$  on the invasiveness of A549 cells. A pipette tip is used to create a wound (in yellow) in a confluent culture (in green). Two marks, drawn with a permanent marker (red and blue line) were used to identify univocally four regions of the wound (red and blue shaded areas) that were monitored over time to determine the closing rate of the wound.

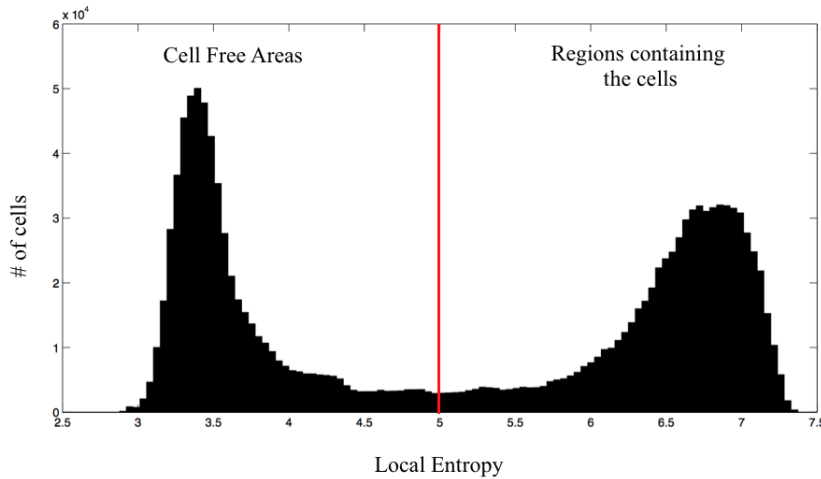


Figure 5.20: The histogram of the local entropy image is clearly bimodal, making the Otsu's method an ideal strategy to segment the wound.

spread. Despite being very simple, this method is very effective in the current application due to the shape of the distribution of the local entropy in scratch wound healing assay images (Figure 5.20). Its definite bimodality allows to easily and correctly identify the two classes.

As detailed in [124] there are four possible invocations for Cell-Invasiv-O-Meter, that differ for the number and the kind of arguments required.

```
[ file , path]= CellInvasivOMeter( folderImages )
```

This is the simplest options, in which all the parameters are left to their default values.

```
[ file , path]= CellInvasivOMeter( folderImages , kernelDim )
```

In this version of Cell-Invasiv-O-Meter invocation the user modifies the size of the neighbourhood used to compute the local entropy. The default value for this parameter is 45 and it is inversely proportional to the resolution of the entropy image (the bigger the kernel, the lower the resolution).

```
[ file , path]= CellInvasivOMeter( folderImages , woundPrevalence )
```

In this third option the second argument is a number between 0 and 1 (default 0.7) that represents the fraction of cell free area that is part

of the wound. Changing this parameter can be useful when the cell culture has a low confluence and thus there are a number of “holes” that don’t belong to the wound and must be excluded.

```
[file , path]=CellInvasivOMeter(folderImages , kernelDim ,  
    woundPrevalence)
```

This alternative combines the two previous versions of the function’s invocation.

At the end of the elaboration a dialog window will appear and the user will be able to provide a name and a location for the .mat file in which the results will be saved.

### Data Analysis

The analysis of the results and the production of the final graphs is executed by another function named `compare-Cell-Invasive-o-Measures` (freely available at [124]).

It loads the .mat file produced by `Cell-Invasiv-o-meter` and computes the average cell free area for every time point and condition. As explained in the previous section the wound area is measured in pixels rather than  $\mu M^2$  to make this function independent of the set-up used and specifically of the objective’s magnification and to avoid having to assign a regular shape to the wound to compute the equivalent area. Furthermore the results of these assays are generally reported as a relative measure, thus removing any dependence on the measurement unit. Successively `compare-Cell-Invasive-o-Measures` normalizes each area with respect to the value measured for the control at  $T=0$ , to remove any bias in the result due to differences in the initial wound size and computes the rate of closing of the wound as the percentage difference between the normalized area measured at each time point and the one obtained at  $T=0$  for the same condition.

Finally it plots the results as a bar graph (Figure 5.3) that shows the cell free area computed for each condition and time point.

### 5.4.2 I-AbACUS

#### Transwell assay

This *in-vitro* assay allows to quantify the invasiveness of a cellular culture in a more realistic setting. It involves in seeding a defined amount of cells in a Matrigel (Corning Life Sciences, Tewksbury, MA, USA) coated transwell insert (Figure 5.5) and using a gradient in the nutrient’s concentration to induce cell migration across the membrane. To obtain the results described in the previous section two ovarian

cancer cell lines, OVCAR4 and A2780, were used. The former was a kind gift of Dr. Michelle Henderson (Children's Cancer Institute, UNSW, Sydney, Australia), while the latter was generously provided by Dr. Elizabeth Roundhill (Children's Cancer Institute, UNSW, Sydney, Australia). They were both cultured in RPMI-1640 media, supplemented with 10 % foetal bovine serum, penicillin/streptomycin and GlutaMAX (Life Technologies, Carlsbad, CA, USA). Cells were grown at  $37^{\circ}\text{C}$  in 5%  $\text{CO}_2$  and were routinely tested negative for mycoplasma contamination.

Each transwell was coated with 50  $\mu\text{L}$  of Matrigel at concentration of 1 mg/mL and at least 4 hours were allowed for the polymerization to occur at  $37^{\circ}\text{C}$  in 5%  $\text{CO}_2$ .

Cells were harvested with trypsin and the culture's density was automatically determined (Countess, Thermo Fisher, Waltham, Massachusetts, USA) to improve seeding's precision.

Migration was evaluated at 5 initial densities for OVCAR4 cells ( $1 \times 10^5$ ,  $3 \times 10^5$ ,  $5 \times 10^5$ ,  $7 \times 10^5$ ,  $9 \times 10^5$  cells/mL) and at  $5 \times 10^5$  cells/mL for A2780s.

Cells were seeded in the transwell and after 48 hours of incubation were fixed with Methanol 100 % and stained with crystal violet. The membranes were removed from the transwell and mounted on a microscope slide. Representative images from different cell densities are shown in Figure 5.14.

Images of 4 different fields were acquired for each membrane with an optical microscope using a 20x magnification.

Each one of the 3 independent OVCAR4 experiments was repeated in triplicates while, for A2780 cells, 2 independent experiment were conducted, each one consisting of 2 replicates.

### Manual counting

The multi-point tool of ImageJ [113] was used to manually count the cells. Out of focus regions and cells overlapping the left and bottom edge of the images were excluded from the analysis. The time required to count the cells in each image was manually recorded with a stopwatch.

### I-AbACUS counting

I-AbACUS is a custom made program developed in Matlab R2016a [81] that was compiled and is freely available at [www.marilisacortes.com](http://www.marilisacortes.com) either as a standalone application or as source code.

The program's GUI guided the analysis of the data here presented that

were obtained with the default segmentation parameters (15, 9, 11, 17, 19). The scale factor, however, was varied depending on the technical specifications of the computer.

The counting time was measured with a stopwatch and includes both the time required by the program to analyse the images and that used to adjust the cell count.

Figure 5.21 summarizes the main steps of the analysis of transwell assay images using I-AbACUS.

After adjusting the settings, the user is asked to provide the details of the experiment as the name and the number of conditions and replicates and the location of the images to analyse. To ensure the correct association between the images, and thus the cell counts, and the corresponding condition and replicate, I-AbACUS requires the images to be organized in a specific folder structure. Specifically the user must create a directory for every tested condition, each containing one folder for every replicate that, in turn, comprise the corresponding images. Furthermore the replicate folders must have a name that ends with “*\_replicateNumber*” where *replicateNumber* is an integer representing the index of the corresponding replicate.

Once the experiment structure has been outlined the analysis proceeds with the segmentation of the images and the classification of the foreground objects. The result of this operation is shown in panel 4 of Figure 5.21, where object recognized as cells are shown in green, while excluded regions are in red. Selecting an area non correctly segmented/classified leads to the opening of a window like the one in panel 5 (Figure 5.21), where the four alternative segmentation for that specific region are presented. If none of them is correct the user can directly input the number of cells. The cell count is automatically updated after each modification and retained within I-AbACUS for the duration of the analysis.

After all the images have been segmented the user is given the option of visualizing the results, both as a table and as an histogram (panel 8 Figure 5.21), and save the results as an excel file (panel 9 Figure 5.21).

### Image Elaboration and Segmentation

In I-AbACUS images are segmented applying the marker controlled watershed [123] to the saturation channel of images coded in the HSV colour space. This algorithm consists in placing metaphorical water sources in the local minima of the gradient image and draw edges when water from different basins meets. To improve segmentation’s precision, the images are pre-elaborated to remove any gradient minimum not associated to a cell. The foreground regions are marked



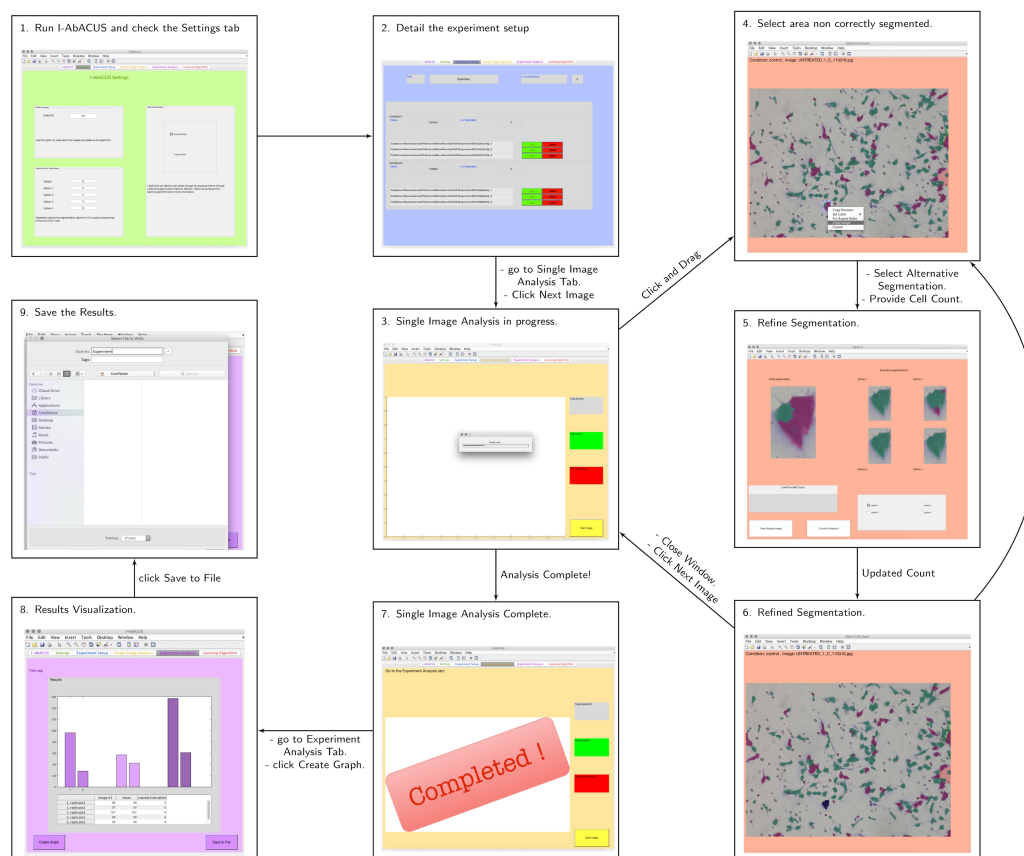


Figure 5.21: Flowchart highlighting the main steps of the analysis of transwell assays with I-AbACUS. After revising the settings (panel 1) the user is required to input the experiment structure, that is specify the number and name of each experimental condition and replicate and location of the images (panel 2). Then each image is segmented and the foreground regions are analysed to determine if they are cells (panel 3). Once all the segmented regions have been analysed the program opens a window like the one in panel 4. Here the user can select parts of the image non-correctly segmented/classified and then select one of the possible alternatives or directly provide the corresponding number of cells. Once all the images have been analysed the results are shown, both in graphical and tabular form (panel 8), and the results can be saved as an excel file (panel 9).

through an opening and closing by reconstruction (Figure 5.6) followed by the removal of the markers with an area smaller than 100 pixels. The kernel used for this operation is a disk of radius equal to one of the segmentation parameters set in the settings panel. This leads to five alternative foreground marker sets corresponding to regions of approximately round shape and size proportional to the corresponding segmentation parameter. The background markers are obtained applying a coarse segmentation to the reconstructed image and then computing the skeleton by influence zone of the background (Figure 5.7). This procedure consists in identifying the ridge lines of the watershed transform computed on the distance transform of the segmented image and ensures a good separation between the foreground and background markers. This approach requires the execution of 5 independent segmentations for each image, that are executed as a parallel process to minimize the elaboration time.

Another procedure applied during the pre-elaboration phase is the background removal, that is obtained segmenting each image with the Otsu's method [116] and assigning to 0 every background pixel.

### Cell Classification

I-AbACUS implements two different strategies to distinguish between cells and non-cells. One is an EF, that was designed to be generally applicable to different types of cells while the other, a trained SVM, allows the user to improve the classification of a specific cell line.

Both methods rely on three morphological characteristics of the cells to make the classification:

- Area: measured in pixels, allows to remove debris and small irregularities.
- Circularity: computed using Equation 5.2 and compared to that of a circle (1). It can prevent pores from being recognized.
- IQR: evaluated on the values of the pixels that belong to each foreground region. Being a measure of dispersion allows to remove debris and out of focus cells.

When the EF is used a foreground region is recognized as a cell if at least 2 of the previously mentioned characteristics assume values within predefined ranges. Specifically the area must be between  $10^2$  and  $10^5$  pixels, the difference between the circularity of the object and that of a circle must be below 1 and the IQR must be above 20 and below 50. Alternatively the user can employ a trained SVM, a classifier that has

been optimized to distinguish a specific kind of cells on the basis of the above mentioned characteristics.

### Learning Algorithm

I-AbACUS integrates a learning algorithm that can be used to train a SVM, a supervised learning model that provided an adequate training set, can be used to improve the recognition of a specific type of cells [126].

The training set used in this study was obtained running the I-AbACUS learning algorithm on a training set composed of 160 images (48 from A2780 cells and 112 from OVCAR4 cells), completely independent from the test set used during I-AbACUS's validation and mostly acquired by Estelle Llsamosas. These images comprise a significant range of densities and were analysed both with the traditional approach and with I-AbACUS obtaining an error that follows a gaussian distribution and is less than 35% in over 90 % of the images (Figures 5.16, 5.17). This trained SVM was chosen as the one that granted the best classification of the OVCAR4 cells in a subset of 9 images, extracted from the dataset used to compare I-AbACUS to the traditional analysis strategy of transwell assays. For this test only the three highest cell densities were considered ( $5 \times 10^5$ ,  $7 \times 10^5$ ,  $9 \times 10^5$  cells/mL) as they are the most difficult to correctly identify, due to the significant number of cells in the image.

The characteristics of the four datasets are presented in Table 5.2, the one named "All cells" includes all the images that were available, while in "All cells, no Delay", were excluded images that were acquired on a day that was different from the one in which the slide was mounted. "OVCAR4" and "OVCAR4, no Delay" are obtained from the previous two datasets removing the images featuring A2780 cells.

The training sets are compared by quantifying the fraction of cells that are correctly recognized in at least one of the five alternative segmentations; the results of this analysis are shown in Table 5.3 and Figure 5.22. Combining different cell lines and including images acquired on a day different than the one in which the slide was mounted, prevents the algorithm from obtaining a classifier that performs better than the random one, but removing either or both of these factors leads to the identification of very similar classifiers that are able to correctly recognize 95 % of the cells, on average.

To obtain the results here presented was used the training set named "All cells, no Delay", because, it was the one featuring the smaller percentage difference between the I-AbACUS and the manual count and it was also able to correctly classify the highest number of cells in the

|                     | Cells         | Images | Mean Difference   | Marker         |
|---------------------|---------------|--------|-------------------|----------------|
| All cells           | OVCAR4, A2780 | 240    | $0.160 \pm 0.282$ | blue circle    |
| All cells, no Delay | OVCAR4, A2780 | 160    | $0.153 \pm 0.277$ | red square     |
| OVCAR4              | OVCAR4        | 192    | $0.174 \pm 0.307$ | green triangle |
| OVCAR4 no delay     | OVCAR4        | 112    | $0.173 \pm 0.312$ | yellow diamond |

Table 5.2: Characteristics of the training sets used in this study. The first one (“All cells”) comprised all the available images, while in the second one were removed all the images that were acquired on a day different from the one in which the corresponding slide was mounted. The third and fourth training sets were obtained from “All cells” and “All cells, no delay” excluding the images of A2780 cells. Here mean difference refers to the average percentage difference between the manual and the I-AbACUS cell counts and marker identify the symbol used to represent the corresponding dataset in Figure 5.22. The dataset chosen to be used in the rest of the analysis is shown in red.

|                    | Initial Density | All cells | All cells, no delay | OVCAR4 | OVCAR4, no delay |
|--------------------|-----------------|-----------|---------------------|--------|------------------|
| 1                  | $5 \times 10^5$ | 0.546     | 0.911               | 0.941  | 0.923            |
| 2                  | $5 \times 10^5$ | 0.721     | 0.977               | 0.891  | 0.932            |
| 3                  | $5 \times 10^5$ | 0.581     | 0.956               | 0.973  | 0.974            |
| 4                  | $7 \times 10^5$ | 0.481     | 0.946               | 0.944  | 0.943            |
| 5                  | $7 \times 10^5$ | 0.480     | 0.980               | 0.981  | 0.987            |
| 6                  | $7 \times 10^5$ | 0.398     | 0.955               | 0.944  | 0.944            |
| 7                  | $9 \times 10^5$ | 0.320     | 0.963               | 0.953  | 0.962            |
| 8                  | $9 \times 10^5$ | 0.391     | 0.968               | 0.944  | 0.971            |
| 9                  | $9 \times 10^5$ | 0.371     | 0.991               | 0.981  | 0.991            |
| mean               |                 | 0.478     | 0.961               | 0.950  | 0.959            |
| standard deviation |                 | 0.118     | 0.022               | 0.026  | 0.023            |

Table 5.3: Table reporting the results of the comparison between the four tested training sets. A subset of 9 images, three for each considered initial cell density ( $5 \times 10^5$ ,  $7 \times 10^5$ ,  $9 \times 10^5$  cells/mL) was analysed with the trained SVMs obtained from the corresponding training sets. Here the percentage of cells that was correctly segmented and classified in at least one of the alternatives computed by I-AbACUS is reported. The SVM that was used for the rest of the analysis (reported in red) has the highest percentage of cells correctly recognized on average.

tested images.

As already remarked in the previous section SVMs are linear classifiers that identify the maximum margin hyperplane, that is the hyperplane that maximizes the distance between the two classes (Figure 5.8, Equation 5.3). To extend the applicability of this strategy to situations in which the two classes are not linearly separable, I-AbACUS implements the kernel method, that consists in using a kernel (in this case a polynomial one) to transform the input space in one of higher dimensionality in which the data are linearly separable and then minimizing Equation

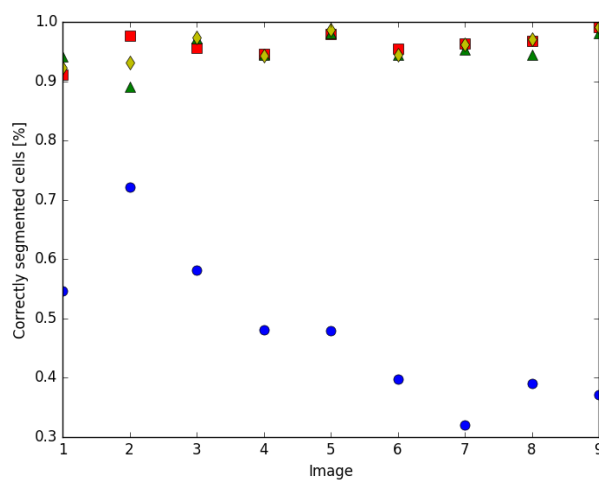


Figure 5.22: Graphical representation of the data shown in Table 5.3. The training set named “All cells” (blue circles) is unable to achieve a classification significantly better than the random one, while the other three training sets achieve substantially equivalent results. The trained SVM used in the following is the one denominated “All cells, no delay” whose results are here shown as red squares.

5.3 in this space (Figure 5.9) .



## Chapter 6

# Conclusions and Perspectives

The main focus of this thesis is the computational representation of a complex biological process involved in phenotypic cell decision making. This process, known as Epithelial to Mesenchymal transition (EMT), is fundamental for wound healing, embryogenesis and stem cell behaviour [7] as it is associated to an increased cell proliferation and augmented migration capabilities. It also presents a relevance in the context of cancer progression specifically in relation to metastasis formation and a poor prognosis.

The framework developed to describe computationally the EMT was designed to be able to integrate the large number of processes that regulate this phenomenon and study their crosstalk. This is realized through the use of simplified representations of the underlying signal transduction networks that do not require a large number of parameters and an extensive experimental characterization of the studied process. Furthermore two descriptions of the same phenomenon at different scales (single cell and population level) are integrated, to evaluate how the behaviour of single cells and the variability within isogenic populations influence the macroscopic behaviour.

The complete description of the model and of its realization is reported in detail together with a preliminary validation of this computational framework, obtained comparing the results of the simulations to microarray data. This analysis, although needing further characterization, highlighted the ability of the model to reproduce the first phases of EMT induction and suggested which aspects of this representation might benefit from a more detailed study. In particular the model was unable to reproduce the behaviour recorded *in-vitro* of two markers (BRK1 and KLK3) that are involved in cell growth, proliferation and immune response. A revision of the model's structure, aimed to improve the representation of these processes, might allow for a more

comprehensive description of the EMT, that would cover all the major steps of this transition.

Once validated, this framework might prove useful to study the EMT in previously untested conditions, to evaluate the effects on this process of different experimental conditions or alternative induction stimuli. This analysis might also inform the experimental study of this process, determining which experiments are the most likely to highlight a phenomenon of interest or a specific behaviour.

Another important characteristic of the presented model is its generality, that is the possibility of using the same procedure to study other biological processes. Indeed most complex phenomena in biology are characterized by large regulatory networks that drive the behaviour of single cells, combined with a tight connection between the behaviour of the population and that of the elements that compose it.

This approach would be very cost-effective. Consider that all the information required for obtaining the results presented in chapters 2 and 3 is freely available online, from specialized databases. This approach improves the usability of this framework but, on the other hand, makes it sensitive to imprecisions and inconsistencies in the data released through these instruments. This framework could thus be used to provide a bird's eye view of the process of interest that would generate a more comprehensive information than more detailed models and either guide the experimental analysis or identify which aspects might require a more detailed representation.

Beside the computational study of EMT, four software tools for the analysis of biological data are presented in this work.

Two of them, detailed in Chapter 4, allow the quantification of the concentration of specific proteins of interest at single-cell level from images acquired with an optical microscope. These protocols were developed to analyse either bacterial or eukaryotic cells and can be used to characterize the expression of specific genes in multiple experimental conditions, both in terms of average protein level and variability within the population.

The possibility, with these tools, of identifying the signal emitted by single cells sets them apart from many gene expression quantification techniques that can only measure population averages. The inability of these techniques to determine phenotypic noise reduces the reliability and the sensitivity of their results, in which interesting behaviours and subtle changes in gene expression might be masked. Thus the presented tools are better suited to inform a computational model, through the identification of its parameters, since their results have an higher level of detail and provide a more accurate representation of the process of



interest

An important requirement for this analysis is a low level of measurement noise and the possibility of recording a reliable signal. For this reason the development of the computational tools presented in chapter 4 was associated with a calibration protocol aimed to compensate for the major aberrations introduced by optical microscopes.

This simple procedure led to the acquisition of data that resulted equivalent to those obtained with a flow cytometer on a wide dynamic range, validating the proposed method and opening the study of phenotypic noise to a larger base of researchers.

Another aspect analysed in the final chapter is the quantification of the invasiveness of a cellular population. This is one of the main macroscopic effects of EMT and a characteristic widely evaluated in cell biology, both to study the behaviour of a population in different experimental conditions and to test new pharmacological treatments. In this regards software tools for the analysis of the data produced from the two most widely used experimental assay in this field were developed.

One of them, studies migration and invasion in a simplified 2D environment, while the other reproduces the phenomenon of interest in a more realistic 3D setting. Both of them, however, are generally analysed using extremely simple and scarcely automatized techniques. This reduces the accuracy of the results and makes the data acquired by different users scarcely comparable.

The use of the computational tools here presented was shown to improve the quality of the findings, while increasing the protocol's standardization. Indeed for the scratch wound healing assay, a simple strategy was devised to be able to univocally identify four regions of the wound and thus remove the uncertainty in the data caused by the selection, at every time point of random fields. Furthermore the quantification of the cell-free area through an automated procedure based on image analysis, significantly improved the accuracy and objectivity of the result.

As for I-AbACUS, the tool developed to analyse transwell assays, an extensive validation procedure, described in a paper currently submitted for publication, demonstrated that it is associated to a reduced intra and inter-operator variability when compared to manual counting and that its performance is robust to changes in cell morphology and density. The widespread application of this computational tool could thus increase the reproducibility and the precision of the resulting data, promoting a more quantitative and precise evaluation of biological phenomena.

Overall the work here presented contributes to the increasingly applied approach that integrates *in-vitro* and *in-silico* techniques when studying biological processes. This increasingly favoured strategy has the potential to both characterize complex natural behaviours, integrating experimental knowledge with *in-silico* inference and make *in-vitro* analysis more efficient and effective integrating additional knowledge in the trial and error process that often characterizes biological research. Furthermore the data analysis tools introduced in the previous chapters promote protocol standardization and transparency, ultimately leading to increased reproducibility and more reliable results.

## 6.1 Future Developments

Both the mathematical model of EMT and the computational tools for the analysis of biological data here described are integrated within a very active research area that aims to provide a more comprehensive representation of complex biological processes through quantitative experimental analyses combined with *in-silico* representations of the same phenomena. Thus several future developments could be realized to improve the quality and usability of the presented projects.

The computational representation of EMT could be modified so as it will be able to reproduce the entire transition. This could be realized updating the structure of the regulatory network used to define the boolean model, so that the behaviour of all the considered markers will be reproduced correctly. Under the hypothesis that the exclusion of one or more elements connected to the regulation of proliferation and immune response is responsible for the inability of the model to proceed with the transition after its first phases, this modification would produce a system able to accurately and completely reproduce the transformation of interest.

This improved computational representation of EMT could then undergo a complete validation that would compare its results with *in-vitro* data acquired *ad-hoc*. This would allow for a more effective comparison, since the experimental conditions could be adjusted to mimic the ones simulated *in-silico*. Furthermore the acquisition of single-cell level data through, for example, immunofluorescence experiments, would make it possible to compare the distributions of expression of each marker. Beside being associated to a more complete validation of the model these data would provide a more precise characterization of the process of interest.

Once validated, the EMT model could be used to evaluate the effect of different experimental conditions on this process. Different cell lines

could be simulated to investigate the modulations of this phenomenon in alternative experimental models and thus determine which one is the most suited for a specific study. This analysis would be particularly helpful when investigating the potential therapeutic effect of EMT reversal and its combination with currently available therapies. Indeed a preliminary *in-silico* screening could significantly reduce the complexity of this study and the time required to complete it, through the identification of the most promising treatments for every experimental model.

Furthermore alternative induction strategies, beside the addition of  $\text{TGF}\beta$ , could be tested, both singularly and in combination, to study their effect on the main steps of EMT. Indeed mechanical stimuli, like compression, or an hypoxic environment have been associated to EMT induction [127, 128] and are widely regarded as better representations of its initiation *in-vivo*. The presented model could thus be used to study the induction of EMT in realistic conditions, that might uncover new and more effective strategies for modulating this transformation, beside expanding the knowledge about this phenomenon in rarely considered conditions.

As for the computational tool presented in Chapter 4, that aims to quantify the level of expression of specific proteins in eukaryotic cells using fluorescent markers, it could be further developed and validated. Additional images, featuring both red and green fluorescent signals, could be tested and these experiment could be executed concurrently on a flow cytometer, to obtain reliable data that could be used as gold standard for the validation of this newly developed tool.

Once its equivalence with well established techniques was determined, this software could be applied to a number of applications in which the evaluation of phenotypic noise could improve the analysis's precision and accuracy. One of such cases is the validation of the computational model of EMT earlier detailed. Indeed the single-cell level data obtained with this technique could be compared to the distributions describing the expression of each marker at different times during the phenotypic transformation, providing enough information to determine the model's ability to describe the considered biological phenomenon.

Furthermore this tool could potentially be used every time the level of expression of specific proteins of interest must be evaluated. While unable to study a large number of markers concurrently, due to the longer time required to process each sample with respect to high throughput techniques, the increased precision and level of detail of the results provided by this method would make it worth using for the

study of the most relevant regulators of the considered process.

Finally I-AbACUS, the software developed in collaboration with Dr. Ford at UNSW, could be further extended to include a number of features that would significantly improve the usability and utility of this tool. One of the possible developments is the de-identification of the images, that could remove the bias, connected to the expected result, that the operator could introduce during the analysis. This upgrade, simply realized presenting the images in random order to the user and removing the indication of experimental condition and replicate, would further improve the objectivity of this analysis method.

Furthermore a framework aimed to facilitate the troubleshooting phase of transwell invasion/migration assays could be integrated within I-AbACUS. This tool could be extremely useful to reduce the time required to identify the most appropriate initial cell density, that generally requires an extensive trial and error optimization. A standard protocol for the identification of this parameter would also increase the assay's repeatability between different laboratories and facilitate the planning of these experiments.

Finally I-AbACUS could be updated to include a training program, able to provide feedbacks and suggestions to inexperienced users. It would consist in the analysis of a set of images, specifically chosen to highlight the main difficulties of the interpretation of these experiments, and in the comparison of the trainee's results with the ones of an experienced user. This tool would provide new users with clear and objective classification criteria that would reduce the time required for the training through an increased efficiency of the learning process.

The development of a structured training framework for the analysis of transwell assay, could also contribute to the repeatability and reliability of the results of this technique, reducing the subjectivity of the classification process and the dependency of the cell counts on the user's experience. Ultimately, this could lead to the development of a fully automated tool for the elaboration of these images, that would improve the quality of the analysis and reduce the time required to complete it. Both these characteristics would significantly upgrade this technique and possibly extend further its use.

In conclusion the presented work is comprised within a very active and exciting research field, that benefits significantly from the integration of knowledge developed in multiple disciplines and the cooperation of researchers with different backgrounds and expertises. Indeed the innovative approaches developed intersecting technical and biological know-how could revolutionize the way biological phenomena are studied, both providing tools for the development of new treatments and

improving our understanding of the mechanisms that drive cellular processes, through accurate, quantitative and reliable data.



# Acknowledgements

As I reach the end of my PhD I cannot help but thinking back at the beginning of this journey, at the excitement at the happiness and at the apprehension I felt three years ago.

While a lot of things have changed in this time, me included, the people I would like to thank and to whom I would like to dedicate this work have not. They are the pillars of my life, the ones that always supported me and pushed me to work harder, dream bigger and think out of the box. First and foremost the two great women of my family, my mother and her mother, my greatest fans, I could never thank you enough for everything that you did and do for me. To Andrea, my partner in crime, the other half of me, I would never have gotten here without you. And of course I could not forget the rest of my family, my father, grandfather, brothers, cousins, aunts and uncles that helped and supported me along the way.

Then I would like to thank all the people that shared this journey with me, Emanuele and Simone, my supervisors, always kind, attentive and competent. Thank you for guiding me during these three years while giving me the space and freedom to find my way in this world. Lucia, Alice and Joseph, my colleagues and friends, thank you for the support, the long talks and the help, and of course the Apice's gang and all the "interesting" lunches we've shared, you've been the comic relief I needed.

Finally thank to Caroline, Raani, Claire and all my Aussie friends, you made me feel at home on the other side of the world and I have to confess a piece of me stayed with you when I left.





# Bibliography

- [1] G Balazsi, A van Oudenaarden, and JJ Collins. Cellular decision-making and biological noise: From microbes to mammals. *Cell*, 2011.
- [2] O Brandman, JE Ferrell, Jr, R Li, and T Meyer. Interlinked fast and slow positive feedback loops drive reliable cell decisions. *Science*, 2005.
- [3] JC Ray and OA Igoshin. Adaptable functionality of transcriptional feedback in bacterial two-component systems. *PLOS Computational Biology*, 2010.
- [4] A Tiwari, G Balazsi, ML Gennaro, and OA Igoshin. The interplay of multiple feedback loops with post-translational kinetics results in bistability of mycobacterial stress response. *Physical Biology*, 2010.
- [5] MB Elowitz, AJ Levine, ED Siggia, and PS Swain. Stochastic gene expression in a single cell. *Science*, 2002.
- [6] G Magyar, A Kun, B Oborny, and JF Stuefer. Importance of plasticity and decision-making strategies for plant resource acquisition in spatio-temporally variable environments. *New Phytology*, 2007.
- [7] S Lamouille, J Xu, and R Derynck. Molecular mechanisms of epithelial-mesenchymal transition. *Nature Reviews Molecular and Cellular Biology*, 2014.
- [8] MA Nieto, RY-J Huang, RA Jackson, and JP Thiery. EMT: 2016. *Cell*, 2016.
- [9] H Acloque, MS Adams, K Fishwick, M Bronner-Fraser, and MA Nieto. Epithelial-mesenchymal transitions: the importance of changing cell state in development and disease. *Journal of Clinical Investigation*, 2009.

- [10] MA Nieto. Epithelial plasticity: a common theme in embryonic and cancer cells. *Science*, 2013.
- [11] JP Thiery, H Acloque, RY Huang, and MA Nieto. Epithelial-mesenchymal transitions in development and disease. *Cell*, 2009.
- [12] J Xu, S Lamouille, and R Derynck. TGF- $\beta$ -induced epithelial to mesenchymal transition. *Cell Research*, 2009.
- [13] MY Maitah, S Ali, A Ahmad, S Gadgeel, and FH Sarkar. Up-regulation of sonic hedgehog contributes to TGF- $\beta$ 1-induced epithelial to mesenchymal transition in NSCLC cells. *PLOS ONE*, 2011.
- [14] R Kalluri and RA Weinberg. The basics of epithelial-mesenchymal transition. *The Journal of Clinical Investigation*, 2009.
- [15] KR Fisher, A Durrans, S Lee, J Sheng, F Li, SF Wong, H Choi, T El Rayes, S Ryu, J Troeger, RF Schwabe, LT Vahdat, NK Altorki, V Mittal, and D Gao. Epithelial to mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. *Nature*, 2015.
- [16] X Zheng, JL Carstens, L Kim, M Scheible, J Kaye, H Sugimoto, CC Wu, VS LeBleu, and R Kalluri. Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. *Nature*, 2015.
- [17] C Kudo-Saito, H Shirako, T Takeuchi, and Y Kawakami. Cancer metastasis is accelerated through immunosuppression during snail-induced emt of cancer cells. *Cancer Cell*, 2009.
- [18] H Beug. Breast cancer stem cells: eradication by different therapies? *Cell*, 2009.
- [19] RY Huang, P Guilford, and JP Thiery. Early events in cell adhesion and polarity during epithelial-mesenchymal transition. *Journal of Cell Science*, 2012.
- [20] JP Thiery and JP Sleeman. Complex networks orchestrate epithelial-mesenchymal transition. *Nature Review Molecular and Cellular Biology*, 2006.
- [21] KN Chua, KL Poon, J Lim, WJ Sim, RY Huang, and JP Thiery. Target cell movement in tumor and cardiovascular diseases based

- on the epithelial-mesenchymal transition concept. *Advancements in Drug Delivery Review*, 2011.
- [22] KS Choudhary, N Rohatgi, S Haildorsson, E Briem, T Gudjonsson, S Gudmundsson, and O Rolfsson. EGFR signal-network reconstruction demonstrates metabolic crosstalk in EMT. *PLOS Computational Biology*, 2016.
- [23] L Jiang, L Xiao, H Sugiura, X Huang, A Ali, M Kuro-o, RJ Deberardinis, and DA Boothman. Metabolic reprogramming during TGF- $\beta$ 1-induced epithelial-to-mesenchymal transition. *Oncogene*, 2015.
- [24] D Mathow, F Chessa, M Rabionet, S Kaden, R Jennemann, R Sandhoff, HJ Gröne, and A Feuerborn. Zeb1 affects epithelial cell adhesion by diverting glycosphingolipid metabolism. *EMBO*, 2015.
- [25] YD Shaul, E Freinkman, WC Comb, JR Cantor, WL Tam, P Thiru, D Kim, N Kanarek, ME Pacold, and Chen WW et al. Dihydropyrimidine accumulation is required for the epithelial-mesenchymal transition. *Cell*, 2014.
- [26] MP Mak, P Tong, L Diao, RJ Cardnell, DL Gibbons, WN William, F Skoulidis, ER Parra, J Rodriguez-Canales, II Wistuba, JV Heymach, JN Weinstein, KR Coombes, J Wang, and L Averett Byers. A patient-derived, pan-cancer EMT signature identifies global molecular alterations and immune target enrichment following epithelial-to-mesenchymal transition. *Clinical Cancer Research*, 2016.
- [27] TCGA website. <https://cancergenome.nih.gov>.
- [28] JK Pauling, AG Christensen, R Batra, N Alcaraz, E Barbosa, MR Larsen, HC Beck, R Leth-Larsen, V Azevedo, HJ Dizel, and J Baumbach. Elucidation of epithelial-mesenchymal transition-related pathways in triple-negative breast cancer cell line model by multi-omics interactome analysis. *Integrative Biology*, 2014.
- [29] P Desai, J Yang, B Tian, H Sun, M Kalita, H Ju, A Paulucci-Holthauzen, Y Zhao, AR Brasier, and RG Sadygov. Mixed-effects model of epithelial-mesenchymal transition reveals rewiring of signaling network. *Cellular Signalling*, 2015.
- [30] X-J Tian, H Zhang, and J Xing. Coupled reversible and irreversible bistable switches underlying TGF- $\beta$ -induced epithelial to mesenchymal transition. *Biophysical Journal*, 2013.

- [31] DT Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 1977.
- [32] SN Steinway, JG Zañudo, W Ding, CB Rountree, DJ Feith, TP Jr Loughran, and R Albert. Network modeling of TGF $\beta$  signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint sonic hedgehog and Wnt pathway activation. *Cancer Research*, 2014.
- [33] SN Steinway, JG Zañudo, PJ Michel, DJ Feith, Loughran TP, and R Albert. Combinatorial interventions inhibit TGF $\beta$ -driven epithelial-to-mesenchymal transition and supports hybrid cellular phenotypes. *NPJ Systems Biology and Applications*, 2015.
- [34] DA Vargas, O Bates, and MH Zaman. Computational model to probe cellular mechanics during epithelial-mesenchymal transition. *Cells Tissues Organs*, 2013.
- [35] A Neagu, V Mironov, I Kosztin, B Barz, M Neagu, RA Moreno-Rodriguez, RR Markwald, and G Forgacs. Computational modeling of epithelial-mesenchymal transformations. *Biosystems*, 2010.
- [36] CH Pratt, R Vadigepalli, P Chakravarthula, GE Gonye, NJ Philp, and GB Grunwald. Transcriptional regulatory network analysis during epithelial-mesenchymal transformation of retinal pigment epithelium. *Molecular Vision*, 2008.
- [37] A Papoulis. *Probability, Random Variables and Stochastic Processes*, chapter Brownian Movement and Markoff Processes. New York, McGraw- Hill, 1984.
- [38] IB Djordjevic. Markov chain-like quantum biological modeling of mutations, aging, and evolution. *Life*, 2015.
- [39] M Shamir, Y Bar-On, R Phillips, and R Milo. Snapshot: Timescales in cell biology. *Cell*, 2016.
- [40] A Flamholz, R Phillips, and R Milo. The quantified cell. *Molecular Biology of the Cell*, 2014.
- [41] MK Doherty, DE Hammond, MJ Clague, SJ Gaskell, and RJ Beynon. Turnover of the human proteome: determination of protein intracellular stability by dynamic silac. *Journal of Proteomic Research*, 2009.
- [42] FM Boisvert, Y Ahmad, M Gierlinski, F Charrière, D Lamond, M A Scott, and AI Lamon. A quantitative spatial proteomics

- analysis of proteome turnover in human cells. *Molecular and Cellular Proteomics*, 2012.
- [43] SC Manolagas. Birth and death of bone cells: basic regulatory mechanisms and implications for the pathogenesis and treatment of osteoporosis. *Endocrinology Review*, 2000.
- [44] Epithelial to mesenchymal transition  $RT^2$  profiler PCR arrays. <https://www.qiagen.com/us/shop/pcr/primer-sets/rt2-profiler-pcr-arrays/?catno={PAHS-090Z}#geneglobe>.
- [45] Kyoto Encyclopedia of Genes and Genomes. <http://www.genome.jp/kegg/>.
- [46] M Kanehisa and S Goto. Kegg: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acid Research*, 2000.
- [47] Kyoto encyclopedia of genes and genomes: Statistics. <http://www.genome.jp/kegg/docs/statistics.html>.
- [48] Meta stability. <https://commons.wikimedia.org/wiki/File:Meta-stability.svg>.
- [49] T Tamura and T Akutsu. An improved algorithm for detecting a singleton attractor in a boolean network consisting of and/or nodes. In *International Conference on Algebraic biology*, 2008.
- [50] A Veliz-Cuba, B Aguilar, F Hinkelmann, and R Laubenbacher. Steady state analysis of boolean molecular network models via model reduction and computational algebra. *Bioinformatics*, 2014.
- [51] N Berntsen and M Ebeling. Detection of attractors of large boolean networks via exhaustive enumeration of appropriate subspaces of the state space. *Bioinformatics*, 2013.
- [52] M Hopfensitz, C Müssel, M Mauchen, and HA Kestler. Attractors in boolean networks: a tutorial. *Computational Statistics*, 2013.
- [53] M Chaves and L Tournier. Predicting the asymptotic dynamics of large biological networks by interconnections of boolean modules. In *50th IEEE Conference on Decision and Control*, 2011.
- [54] NCBI gene database. <https://www.ncbi.nlm.nih.gov/gene>.
- [55] The human protein atlas. <http://www.proteinatlas.org>.

- [56] Dijkstra's algorithm pseudocode. [https://en.wikipedia.org/wiki/Dijkstra's\\_algorithm](https://en.wikipedia.org/wiki/Dijkstra's_algorithm).
- [57] MA Sartor, V Mahavisno, VG Keshamouni, J Cavalcoli, Z Wright, A Karnovsky, R Kuick, HV Jagadish, B Mirel, T Weymouth, B Athey, and GS Omenn. ConceptGen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics*, 2009.
- [58] NCBI Gene Expression Omnibus. <https://www.ncbi.nlm.nih.gov/geo/>.
- [59] S Kurita, E Gunji, K Ohashi, and K Mizuno. Actin filaments-stabilizing and bundling activities of cofilin-phosphatase Slingshot-1. *Genes to Cells*, 2007.
- [60] SW Han and J Roman. Fibronectin induces cell proliferation and inhibits apoptosis in human bronchial epithelial cells: pro-oncogenic effects mediated by PI3-kinase and NF-kappa B. *Oncogene*, 2006.
- [61] N Shao, Y Chai, Cui JQ, N Wang, K Aysola, ES Reddy, and Rao VN. Induction of apoptosis by Elk-1 and deltaElk-1 proteins. *Oncogene*, 1998.
- [62] Y Kawasaki, R Sato, and T Akiyama. Mutated APC and Asef are involved in migration of colorectal tumour cells. *Nature Cell Biology*, 2003.
- [63] N Sakai and K Shibata. Functional characterization of novel tumor suppressor protein SAV1 in cancer cell proliferation and epithelial-mesenchymal transition. *The FASEB Journal*, 2003.
- [64] TGF- $\beta$  signaling pathway. [http://www.genome.jp/kegg-bin/show\\_pathway?hsa04350+8454](http://www.genome.jp/kegg-bin/show_pathway?hsa04350+8454).
- [65] TU Barbie, G Axele, AR Aref, S Li, Z Zhu, X Zhang, Y Imaura, TC Thai, Y Huang, M Bowden, J Herndon, TJ Cohoon, T Fleming, P Tamayo, JP Mesirov, S Ogino, KK Wong, MJ Ellis, WC Hahn, Barbie DA, and WE Gillanders. Targeting an IKBKE cytokine network impairs triple-negative breast cancer growth. *The Journal of Clinical Investigation*, 2014.
- [66] SK Gopal, DW Greening, EG Hanssen, HJ Zhu, RJ Simpson, and RA Mathias. Oncogenic epithelial cell-derived exosomes containing Rac1 and PAK2 induce angiogenesis in recipient endothelial cells. *Oncotarget*, 2016.

- [67] W Xiong and JT Parsons. Induction of apoptosis after expression of PYK2, a tyrosine kinase structurally related to focal adhesion kinase. *Journal of Cellular Biology*, 1997.
- [68] S Zrihan-Licht, Y Fu, J Settleman, K Schinkmann, L Shaw, I Keydar, S Avraham, and H Avraham. RAFTK/Pyk2 tyrosine kinase mediates the association of p190 rhoGAP with RasGAP and is involved in breast cancer cell invasion. *Oncogene*, 2000.
- [69] K Mavridis, M Avgeris, and A Scorilas. Targeting kallikrein-related peptidases in prostate cancer. *Expert Opinion on Therapeutic Targets*, 2014.
- [70] Pathways in cancer, kegg. [http://www.genome.jp/kegg-bin/show\\_pathway?hsa05200+369](http://www.genome.jp/kegg-bin/show_pathway?hsa05200+369).
- [71] Adherens junctions, kegg. [http://www.genome.jp/kegg-bin/show\\_pathway?hsa04520](http://www.genome.jp/kegg-bin/show_pathway?hsa04520).
- [72] C Guo, S Liu, J Wang, MZ Sun, and FT Greenaway. ACTB in cancer. *Clinica Chimica Acta*, 2012.
- [73] G Pothoulakis, F Ceroni, B Reeve, and T Ellis. The spinach RNA aptamer as a characterization tool for synthetic biology. *ACS Synthetic Biology*, 2014.
- [74] F Ceroni, S Furini, Stefan, A A, Hochkoeppler, and E Giordano. A synthetic post transcriptional controller to explore the modular design of gene circuits. *ACS Synthetic Biology*, 2012.
- [75] Laser-induced fluorescence. [https://commons.wikimedia.org/wiki/File:Laser-induced\\_fluorescence.png](https://commons.wikimedia.org/wiki/File:Laser-induced_fluorescence.png).
- [76] Lucia Bandiera, Simone Furini, and Emanuele Giordano. Phenotypic variability in synthetic biology applications: dealing with noise in microbial gene expression. *Frontiers in Microbiology*, 2016.
- [77] Murat Acar, Jerome T. Mettetal, and Alexander van Oudenaarden. Stochastic switching as a survival strategy in fluctuating environments. *Nature Genetics*, 2008.
- [78] J.C Locke and M.B. Elowitz. Using movies to analyse gene circuits dynamics in bacteria using fluorescence time-lapse microscopy. *Nature Protocols*, 2009.

- [79] J.W Young, J.C. Locke, A. Altinok, N. Rosenfeld, T. Bacarian, P.S. Swain, E Mjolsness, and M.B. Elowitz. Measuring single-cell gene expression dynamics in bacteria using time-lapse microscopy. *Nature Protocols*, 2011.
- [80] S.E. Cohen, M.L. Erb, J. Selimkhanov, G. Dong, J Hasty, G Pogliano, and S.S. Golden. Dynamic localization of cyanobacterial circadian clock proteins. *Current Biology*, 2014.
- [81] Matlab, The MathWorks, Inc., Natick, Massachusetts, United States.
- [82] Python software foundation, python language reference, version 2.7. <http://www.python.org>.
- [83] L Bandiera, P Pasini, L Pasotti, S Zucca, G Mazzini, P Magni, E Giordano, and S Furini. Experimental measurement and mathematical modeling of biological noise arising from transcriptional and translational regulation of basic synthetic gene circuits. *Journal of Theoretical Biology*, 2016.
- [84] M Cortesi, L Bandiera, A Pasini, A Bevilacqua, A Gherardi, S Furini, and E Giordano. Reliable measurement of *E.coli* single cell fluorescence distribution using a standard microscope set-up. *Journal of Biological Engineering*, 2017.
- [85] Molecular Expressions. <http://micro.magnet.fsu.edu/primer/java/fluorescence/photobleaching/>.
- [86] Nathalie B. Vicente, Javier E. Diaz Zamboni, Javier F. Adur, Enrique V. Paravani, and Victor H. Casco. Photobleaching correction in fluorescence microscopy images. In *16th Argentine Bio-engineering Congress and 5th Conference of Clinical Engineering*, 2007.
- [87] T Mitsunaga and S.K. Nayar. Radiometric self calibration. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1999.
- [88] A. Bevilacqua, A. Gherardi, and L. Carrozza. A robust approach to reconstruct experimentally the camera response function. In *Proceedings of the IEEE Image Processing Theory, Tools and Applications*, 2008.
- [89] Cave lab, radiometric camera calibration. [http://www.cs.columbia.edu/CAVE/projects/rad\\_cal/](http://www.cs.columbia.edu/CAVE/projects/rad_cal/).



- [90] P.C. Hansen. regtool, matlab toolbox for analysis and solution of discrete ill-posed problems. <http://www.mathworks.com/matlabcentral/regtools>.
- [91] D Marr and E Hildreth. Theory of edge detection. In *Proceedings of the Royal Society of London, Series B. Biological Sciences*, pages 187–217, 1980.
- [92] L Gerosa, K Kochanowski, M Heinemann, and U Sauer. Dissecting specific and global transcriptional regulation of bacterial gene expression. *Molecular System Biology*, 2013.
- [93] V Shahrezaei and S Marguerat. Connecting growth with gene expression: of noise and numbers. *Current Opinion in Microbiology*, 2015.
- [94] Delta optical thin films, what is optical density? <http://www.deltaopticalthinfilm.com/optical-density/>.
- [95] N Navin, A Krasnitz, L Rodgers, K Cook, J Meth, J Kendall, M Riggs, Y Eberling, J Troge, V Grubor, D Levy, P Lundin, S Måner, A Zetterberg, J Hicks, and M Wigler. Inferring tumor progression from genomic heterogeneity. *Genome Research*, 2010.
- [96] DA Lawson, NR Bhakta, K Kessenbrock, KD Prummel, Y Yu, K Takai, A Zhou, H Eyob, S Balakrishnan, CY Wang, AG Yaswen, and Z Werb. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature*, 2015.
- [97] M Krönig, M Walter, V Drendel, M Werner, CA Jilg, AS Richter, R Backofen, D McGarry, M Follo, W Schultze-Seemann, and R Schüle. Cell type specific gene expression analysis of prostate needle biopsies resolves tumor tissue heterogeneity. *Oncotarget*, 2014.
- [98] CKY Ng, LG Martellotto, A Gauthier, HC Wen, S Piscuoglio, RS Lim, CF Cowell, PM Wilkerson, P Wai, DN Rodrigues, L Arnould, FC Geyer, SE Bromberg, M Lacroix-Trili, F Penault-Llorca, S Giard, X Sastre-Garau, R Natrajan, L Norton, PH Cottu, B Weigelt, A Vincent-Salomon, and JS Reis-Filho. Intra-tumor heterogeneity and alternative driver genetic alterations in breast cancers with heterogeneous *HER2* gene amplification. *Genome Biology*, 2015.

- [99] J Wu and ES Tzanakakis. Deconstructing stem cell population heterogeneity: Single-cell analysis and modeling approaches. *Biotechnology Advances*, 2013.
- [100] Y Hasegawa, D Taylor, DA Ovchinnikov, EJ Wolvetang, L de Torrenté, and JC Mar. Variability in gene expression identifies transcriptional regulators of early human embryonic development. *PLOS genetics*, 2015.
- [101] R Bahar, CH Hartmann, KA Rodriguez, AD Denny, RA Busuttil, ME Dolle, RB Calder, GB Chisholm, BH Polleck, CA Klein, and J Vijg. Increased cell to cell variation in gene expression in ageing mouse heart. *Nature*, 2006.
- [102] A Ståhlberg, Kubista M, and P Åman. Single-cell gene-expression profiling and its potential diagnostic applications. *Expert Review Molecular Diagnostics*, 2011.
- [103] NE Navin. The first five years of single-cell cancer genomics and beyond. *Genome Research*, 2015.
- [104] C Gawad, W Koh, and SR Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 2016.
- [105] A Raj and A van Oudenaarden. Nature, nurture or chance: Stochastic gene expression and its consequences. *Cell*, 2008.
- [106] Dapi. <https://en.wikipedia.org/wiki/DAPI>.
- [107] S Beucher and C Lantuejoul. Use of watershed in contour detection. In *International Workshop on image processing: Real-time Edge and Motion detection/estimation*, 1979.
- [108] American Type Culture Collection US, MCF7 (ATCC HTB-22). <https://www.lgcstandards-atcc.org/Products/All/HTB-22.aspx>.
- [109] mcbeng.it. <http://www.mcbeng.it/en/>.
- [110] P Singh and AK Garg. Morphology based non uniform background removal for particle analysis: A comparative study. *International Journal of Computing and Corporate Research*, 2011.
- [111] Chun Chi Liang, Ann Y Park, and Jun Lin Guan. In vitro scratch assay: a convenient and inexpensive method for analysis of cell migration in vitro. *Nature Protocols*, 2(2):329–333, February 2007.

- [112] James E.N. Jonkman, Judith A. Cathcart, Feng Xu, Miria E. Bartolini, Jennifer E. Amon, Katarzyna M. Stevens, and Pina Colarusso. An introduction to wound healing assay using live-cell microscopy. *Cell Adhesion and Migration*, 2014.
- [113] CA Schneider, WS Rasband, and KW Eliceiri. NIH image to Imagej: 25 years of image analysis. *Nature Methods*, 2012.
- [114] Tobias Geback, Martin Michael Peter Schultz, Petros Koumoutsakos, and Michael Detmar. TScratch: a novel and simple software tool for automated analysis of monolayer wound healing assays. *BioTechniques*, 2009.
- [115] C.E Shannon. A mathematical theory of communication. *The Bell System technical Journal*, 1948.
- [116] N Otsu. A threshold selection method for gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 1979.
- [117] CE Henry, E Llsamosas, A Djordjevic, NF Hacker, and CE Ford. Migration and invasion is inhibited by silencing ROR1 and ROR2 in chemoresistant ovarian cancer. *Oncogenesis*, 2016.
- [118] S Seetharaman, E Flemyng, J Shen, MR Conte, and AJ Ridley. The RNA-binding protein LARP4 regulates cancer cell migration and invasion. *Cytoskeleton*, 2016.
- [119] XX Yu, Z Hu, X Shen, LY Dong, WZ Zhou, and WH Hu. IL-33 promotes gastric cancer cell invasion and migration via ST2-ERK1/2 pathway. *Digestive Diseases and Sciences*, 2015.
- [120] KC Chiang, TS Yeh, RC Wu, JHS Pang, CT Cheng, SY Wang, HH Juang, and CN Yeh. Lipocalin 2 (LCN2) is a promising target for cholangiocarcinoma treatment and bile LCN2 level is a potential cholangiocarcinoma diagnostic marker. *Scientific Reports*, 2016.
- [121] P Dong, M Kaneuchi, H Watari, J Hamada, S Sudo, J Ju, and N Sakuragi. Micro-RNA-194 inhibits epithelial to mesenchymal transition of endothelial cancer cells by targeting oncogene BMI-1. *Molecular Cancer*, 2011.
- [122] M-H Yang, Z-Y Hu, C Xu, L-Y Xie, X-Y Wang, S-Y Chen, and Z-G Li. MALAT1 promotes colorectal cancer proliferation/migration/invasion via prka kinase anchor protein 9. *Biochimica and Biophysica Acta (BBA)- Molecular Basis of Diseases*, 2015.

- [123] F Meyer and S Beucher. Morphological segmentation. *Journal of Visual Communication and Image Reconstruction*, 1990.
- [124] M Cortesi. Cell-invasiv-o-meter. <https://github.com/MarilisaCortesi/Cell-Invasiv-o-Meter>.
- [125] Matt Smith. imoverlay. <http://it.mathworks.com/matlabcentral/fileexchange/42904-imoverlay>.
- [126] C Cortes and V Vapnik. Support vector networks. *Machine Learning*, 1995.
- [127] JM Tse, G Chen, JA Tyrrell, SA Wilcox-Adelmann, Y Boucher, RK Jain, and LL Munn. Mechanical compression drives cancer cells toward invasive phenotype. *PNAS*, 2012.
- [128] D Ribatti. Epithelial-mesenchymal transition in morphogenesis, cancer progression and morphogenesis. *Experimental Cancer Research*, 2017.